

This Page Is Inserted by IFW Operations  
and is not a part of the Official Record

## **BEST AVAILABLE IMAGES**

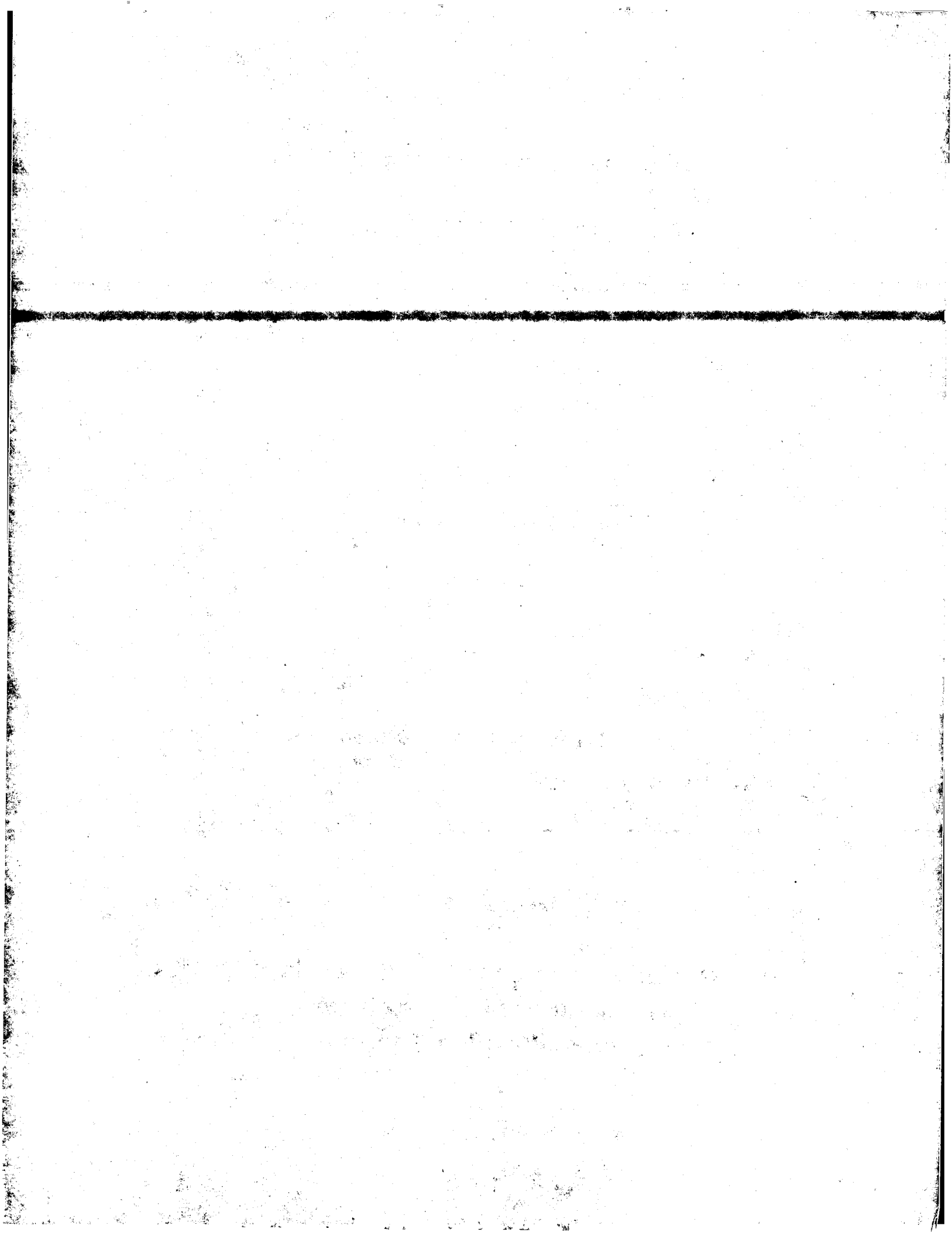
Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

**IMAGES ARE BEST AVAILABLE COPY.**

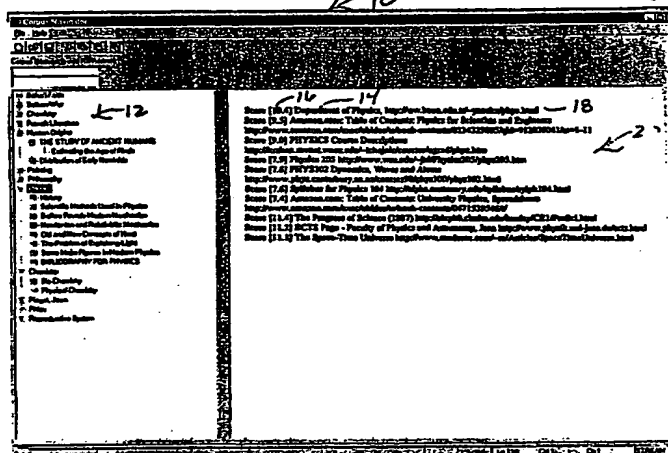
**As rescanning documents *will not* correct images,  
please do not report the images to the  
Image Problem Mailbox.**





## INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification <sup>7</sup> : <b>G06F 17/30</b>		A2	(11) International Publication Number: <b>WO 00/62198</b>
			(43) International Publication Date: 19 October 2000 (19.10.00)
(21) International Application Number: <b>PCT/US00/10003</b> (22) International Filing Date: <b>13 April 2000 (13.04.00)</b> (30) Priority Data: 60/129,103                      13 April 1999 (13.04.99)                      US (71) Applicant: <b>INDRAWEB.COM, INC. [US/US]; P.O. Box 323, Concord, MA 01742 (US).</b> (72) Inventors: <b>KON, Henry; 91 Bow Street, Lexington, MA 02420 (US). BURCH, George; P.O. Box 323, Concord, MA 01742 (US).</b> (74) Agents: <b>KELLY, Edward, J. et al.; Foley, Hoag &amp; Eliot, LLP, One Post Office Square, Boston, MA 02109 (US).</b>		(81) Designated States: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CR, CU, CZ, DE, DK, DM, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).  Published <i>Without international search report and to be republished upon receipt of that report.</i>	

(54) Title: **SYSTEMS AND METHODS FOR EMPLOYING AN ORTHOGONAL CORPUS FOR DOCUMENT INDEXING**

## (57) Abstract

The invention provides for indexing and cataloging of content on the Internet, as well as from other stores of information, may be performed by applying a process that employs an orthogonal corpus, or corpora, of information, such as an Encyclopedia. To this end, the processes described herein identify the topics discussed within the corpus. The process also identifies within the corpus a set of keywords that are relevant to the topics presented in the corpus. The keywords associated with a topic may be employed to identify documents stored in another database that are related to the topic. A graphical representation of the index of topics found in the corpus may then be generated, with individual topics operating as links to these related documents. Thus, a user interested in reviewing content in the corpus related to a certain topic, may also activate a link in the graphical representation of the index to access other documents that have been identified as related to the topic of interest to the user.

**FOR THE PURPOSES OF INFORMATION ONLY**

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav	TM	Turkmenistan
BF	Burkina Faso	GR	Greece		Republic of Macedonia	TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's	NZ	New Zealand		
CM	Cameroon		Republic of Korea	PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakhstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		
EE	Estonia	LR	Liberia	SG	Singapore		

# SYSTEMS AND METHODS FOR EMPLOYING AN ORTHOGONAL CORPUS FOR DOCUMENT INDEXING

## Field of the Invention

The invention provides systems and methods for organizing a collection of information, such as a database of documents, and more particularly, to systems and methods for classifying information with respect to a set of independent concepts, and applications of the classified information.

## Background of the Invention

The computer-based storage of information has become the norm for our knowledge-based society, and today huge databases of computer-stored information are available to researchers. For example, through the World Wide Web service, the Internet has become an enormous repository of readily accessible information. However, the sheer volume of information maintained on the World Wide Web becomes problematic when someone wishes to identify those resources, or documents that are maintained on the Internet and that are related to a certain topic. The problem is that a simple search request for information stored on the Internet and related to a certain topic returns an enormous volume of information, most of which is extraneous and irrelevant. This problem arises in part from the difficulty that existing Internet retrieval systems have with prioritizing and categorizing documents stored on the World Wide Web. Today, most systems for organizing content on the web work by searching through a document stored on the web, and associating with the document a set of key words that represent the information contained on the reviewed pages. Accordingly, documents can be organized according to the keywords that appear in the document.

5           However, lists of keyword, although helpful, lack an identifiable hierarchy of  
concept and therefore fail to provide a topic-based system for organization and  
document retrieval. Thus, keyword lists are less powerful tools for document research  
than topic-based hierarchal systems of document organization, such as tables of content,  
or topic indexes. Today there is a need for a "table of content" for the World Wide Web.  
10       Such a table of content would provide a more powerful metaphor for using the Web as a  
research tool and would provide a flexible platform on which other research tools could  
be built. As such, there is a need for improved systems for organizing information  
stored in a computer database, such as the World Wide Web.

### 15       Summary of the Invention

          The systems and methods described herein provide systems for document  
indexing and scoring of content on a computer database, such as the World Wide Web.  
The systems generally include an orthogonal corpus that may comprise a collection of  
blocks of text, and that may be employed to index and score textual information for  
20       applications in retrieving, classifying, or browsing over a set of documents.

          An orthogonal corpus, as the term is employed herein, may be understood to  
encompass, without being limited to, any collection of blocks of text that are outlined or  
referenced by a table of contents, topic index, chapter heading or other topical indicia  
where each topic either stands alone or is an identified subpart (subtopic) of another  
25       topic, forming a tree of topics and their descendant subtopics. Encyclopedias, text and  
reference books, periodicals, web sites, dictionaries, thesauri, the library of congress, the  
Dewey decimal system, and glossaries are examples of, surrogates for, or extenders of  
orthogonal corpora. A set of topics is understood as orthogonal in the sense that  
substantially every member topic (e.g., chapter or article) covers a different concept or  
30       substantially different concept than any other topic under the same ancestor topic in the  
tree. A topic in one practice may be assumed to include or not include, its ancestor  
topics or underlying subtopics.

          The orthogonal construction, or decomposition, of a corpus provides for  
algorithmic identification of keywords in each topic, which distinguish it from its  
35       sibling, cousin, ancestor, or descendent topics. Keywords may be employed to

5 numerically score over an underlying pool of documents. Alternatively, if more practical for a given application, such as when working with a large document collection, rather than individually score all documents in the collection against the keywords, then a set of search words may be generated to identify a subset of candidate documents for scoring.

10 Parameters employed during the scoring process may relate to the identification of keywords and their refinement into a set of search words, with weightings for associated document or sub-document scoring. The scoring, search term, and keyword methods may employ conventional information retrieval techniques including the use of synonyms, stemming, frequency, proximity, stop words, hyponyms, and synonyms.

15 For purposes of clarity certain terms will now be described, although the understandings set forth are not to be understood as limiting and are only provided for purposes of achieving clarity by way of providing examples. The term "word" as employed herein may be understood to encompass a lexical type found in a common or specialty dictionary of any language. The term "phrase" as employed herein may be understood to encompass any sequence of one or more words. Heretofore for simplicity we use "Word" to mean "Word or Phrase." The term "synonym group" shall be understood to encompass a set of words which may be used as alternates for a given word. Each word in a synonym group has a similar or identical meaning. The term "topic" shall be understood to encompass textual content typically having a title, having corresponding text, concerning a single topic, or covering a set or tree of related subtopics. The term "subtopic" shall be understood to encompass a block of text within a Topic. Typically, a subtopic may be identified by its subtopic header or other outline indicator. In different calculation contexts, subtopics may or may not be included as part of parent topics. The term "word\_count" maybe understood as an integer count of the number of times a word or a word in its synonym group occurs in a given topic or text area, potentially including text in the title and headers and any text elements in that text. The term "word\_frequency" may be understood to encompass the word count in a text area divided by the number of words total in that text. A word\_map is a representation of textual content within a text area that is more precise than a word\_count. A word\_map may describe a word's relative location in the text, its linguistic type or contexts of use,

20  
25  
30  
35

5 its prominence indicators such as use in a title or highlighting fonts. The term  
orthogonal corpus may be understood to encompass a collection of topically organized  
information referenced by a table of contents and/or index, where each topic is clearly  
10 identified as a SubTopic of a topic or else stands alone. Nodes in the table of contents  
tree may represent topics. The information may be understood as orthogonal in the  
sense that a stand-alone topic (e.g., chapter or article) covers a substantially different  
concept than any other stand-alone topic, and any subtopic expresses a substantially  
different concept from any other subtopic within the same parent topic. The term  
15 document may be understood to encompass formatted textual content with topic  
beginnings, endings, and marked hierarchy. A document may contain one or more  
topics and may include subtopics. A corpus may include one or more documents. The  
relationship between documents and topics is not mandated, though in some  
embodiments each document to represent one top-level topic along with its subtopics.  
20 The term "discovered document" may be understood to encompass a document (or a set  
of documents such as a web site or portion of a web site) which is being scored. Scoring  
of a discovered document may be relative to one or more corpus documents or corpus  
topics. In one practice, scoring measures the degree of topical relevance to the corpus  
topics. The discovered document will often be a member of a search result set.

25 More particularly, the systems and methods described herein include methods for  
processing a body of reference material to generate a directory for accessing information  
from a database. These methods may comprise processing the body of reference  
material to identify a hierarchical organization of a plurality of topics. Additionally, the  
processes may include the step of associating with at least one of the topics a portion of  
30 the reference material and processing the assigned portion of reference material to  
generate a plurality of search keys representative of search strings for selecting  
information from the database. The process may then apply the search keys to the  
database to retrieve information from the database and may create an association  
between the at least one topic and the information retrieved from the database.

35 In an optional step, the methods described herein may create a graphical interface  
that is representative of the identified hierarchical organization of a plurality of topics



5 for allowing a user to access information retrieved from the database and having an association with the topic. Accordingly, the user may be provided with a graphical interface that allows the user to activate, typically by clicking with a mouse, a graphical representation of a topic to identify a set of links to content, such as web pages that are associated with the topic selected by the user.

10 In one practice, processing the body of reference material includes processing a body of reference material that has been selected from the group consisting of an encyclopedia, a dictionary, a text book, a novel, a newspaper, or a website. Processing the material may include identifying a hierarchical organization of a plurality of substantially orthogonal topics. This may include identifying a table of contents for the  
15 body of reference material, identifying an index for the reference material, identifying chapter or subchapter headings within the reference table, identifying definition entries within a dictionary, and other similar operations that identify different topics that occur within the reference material.

20 Optionally, when processing a body of reference material, the process may normalize the identified hierarchical organization of the plurality of topics.

In one practice when processing the assigned text, the process includes a step of generating a word map that is representative of a statistical analysis of the words contained in the assigned text. Generating the word map may include performing a word count process for determining word frequency of a word within the assigned text  
25 and for employing the word frequency for determining the relevance of a word to the associated topic. Processing the assigned text for different topics may also include a step of identifying a set of key words that have an associated measure intradocument orthogonality.

30 In an optional step, processing the assigned text may include identifying a set of synonyms for extending the search keys. Further, a subset of search keys may be selected that have a predetermined measure of correlation to the topic. The search keys may be applied to the database such as through an Internet search engine, to discover documents that are related to the search keys. Optionally, the Internet search engine may be a meta-search engine.

5           Once documents have been discovered from the database that are related to the search keys, documents may be further processed to determine their relationship to the topics associated with the search keys.

          In the methods described herein creating an association between the at least one topic and the information retrieved from the database may include capturing a location  
10           pointer that is associated with the information retrieved from the database. Creating that association may include generating a data structure for the topic which allows storing location pointers that are associated with information retrieved from the database.

          In another aspect, the systems and methods described herein include systems for organizing a collection of documents. Such systems may comprise an orthogonal corpus  
15           of information that is arranged according to an index of topics, a keyword generator for generating a set of keywords representative of a documents associated with a topic in the index of topics, a scoring system for processing documents within the collection of documents to associate with at least a portion of the documents a score representative of the document's association to a particular topic, and a graphical representation for  
20           depicting at least a portion of the index of topics and having respective portions of the graphical representation linked to documents associated with a respective portion of the index of topics.

          In a further embodiment the systems of the invention may include systems for extending the content of the document. These systems can include a parser for selecting  
25           terms within the document to be extended, an orthogonal corpus of information arranged according to an index of topics, a keyword generator for generating a set of key words representative of a documents association with a topic in the index of topics, and a linking system for processing the documents within a collection of documents, to associate with at least a portion of the document, a score representative of the documents  
30           association to a particular topic and for providing the first document with links to the collection of documents for extending the content of that document.

          Accordingly, the systems and methods described herein may leverage the electronically stored content of the World Wide Web in an intelligent and meaningful way, to provide a database of content organized under an orthogonal and hierarchical  
35           index of topics and subtopics.

5           Other objects of the invention will, in part, be obvious, and, in part, be shown from the following description of the systems and methods shown herein.

### **Brief Description of the Drawings**

10           The foregoing and other objects and advantages of the invention will be appreciated more fully from the following further description thereof, with reference to the accompanying drawing wherein;

          Figure 1 depicts a screen shot of a portion of a orthogonal corpus and a set of documents and the scores associated with those documents.

15           Figures 2A through 2B depict a dataflow diagram of one process according to the invention for processing a body of reference material for organizing a collection of documents according to a hierarchical arrangement of topics provided by the reference material;

          Figure 3 depicts one flow chart diagram of a process according to the invention;

          Figure 4 depicts one system according to the invention; and

20           Figures 5-9 depict a further practice according to the invention organizing content according to indices generated from a plurality of references.

### **Detailed Description of the Illustrated Embodiments**

25           To provide an overall understanding of the invention, certain illustrative embodiments will now be described. However, it will be understood by one of ordinary skill in the art that the systems and methods described herein can be adapted and modified to provide systems and methods suitable for other applications and that other additions and modifications can be made to the illustrated embodiments without departing from the scope hereof.

35           It is a realization of the invention that the indexing and cataloging of content on the Internet, as well as from other stores of information, may be performed by applying a process that employs an orthogonal corpus, or corpora, of information, such as an Encyclopedia. To this end, the processes described herein identify the topics discussed

5 within the corpus. The process also identifies within the corpus a set of keywords that are relevant to the topics presented in the corpus. The keywords associated with a topic may be employed to identify documents stored in another database that are related to the topic. A graphical representation of the index of topics found in the corpus may then be generated, with individual topics operating as links to these related documents. Thus, a  
10 user interested in reviewing content in the corpus related to a certain topic, may also activate a link in the graphical representation of the index to access other documents that have been identified as related to the topic of interest to the user.

Turning to Fig. 1, there is depicted a graphical user interface 10 of the type  
15 created and employed by systems according to the invention. The graphical user interface 10 represents a topic index 12, a portion of which is shown in this illustration. The topic index 12 may be a graphical representation of the table of contents of an encyclopedia, or other corpus. A user may employ the graphic interface 10 to access information that relates to the different topics listed in the index 12. Additionally, the  
20 depicted index 12 includes topics and subtopics, including subtopics of the same ancestor topic. For example, in Fig. 1, the topic Human Origins is the ancestor topic for the subtopics, The Study of Ancient Human and the Distribution of Early Hominids. A topic, or a subtopic, may be understood to include, optionally, its ancestor topics or underlying subtopics.

25 The graphical representation of the index 12 may include a hypertext link, or other linking mechanism, for each topic or subtopic in the index 12. For example, the user may activate the links, as depicted by the highlighted topic PHYSICS in Fig. 1, to retrieve a group of documents having content that is associated with the selected topic. As further depicted by Fig. 1, the system 10 may provide a display 20 such that for a  
30 selected topic or a subtopic, such as the selected topic Physics, a document 18, or a plurality of documents 18, may be presented to the user as documents associated with the topic. In the depicted embodiment, a pointer to the document, such as the title and URL 14 may be presented to the user. Additionally, an associated numerical score 16,  
35 that represents that document's association to the topic may also be presented. The

5 development of such scores 16 will be described in more detail hereinafter. Optionally,  
all the documents associated with a topic may be displayed in a window 20 of the  
system 10.

Turning now to Figures 2A and 2B, dataflow diagrams are presented that  
10 illustrate one process for creating a graphical interface, such as the interface 10 of Fig. 1.  
Specifically, Figures 2A and 2B depict a process 30 wherein a corpus, such as an  
existing published book of reference material, is processed by an orthogonal corpus  
indexing (OCI) process that extracts content signatures and topic indices from the  
corpus' content. The depicted process employs the content signatures to generate search  
15 strings for search engines to identify content associated with topics described in the  
corpus. The retrieved or discovered documents may be examined for content relevance  
and the relevant documents may be associated with topics presented in the orthogonal  
index of the corpus. Optionally, site attributes such as document type, timeliness, source  
and other such attributes may also be identified and employed to select relevant websites  
20 that may be associated with a topic in the index of the orthogonal corpus.

More specifically Fig. 2A depicts that the process 30 operates on a corpus 32 that  
may be input to the index generator 34. The index generator 34 may generate an index  
for the corpus 32 and this index may be provided to the keyword generator 48. The  
25 keyword generator 48 may produce a set of key words 52 and may be associated with  
the index 40. The process 30 continues in Figure 2B which shows the index 40 in the  
search keys 52 being applied to a search engine 54. The search engine 54 discovers  
documents from a database of content, or from a collection of databases of content 58 to  
thereby create an association between at least one of the topics of the index 40 and the  
30 information retrieved from the database 58.

The depicted corpus 32 may be any collection of information and may include,  
but is not limited to, encyclopedias, text books, dictionaries, thesauruses, atlases, maps,  
and other reference material. In one embodiment, the corpus 32 may be a published  
35 book that may be turned into or stored in an electronic format such as a conventional

5 computer datafile of text information. The corpus 32, preferably in an electronic format, may be provided to the index generator 34. The index generator 34 may process the corpus 32 to identify a hierarchical organization of a plurality of topics that appear within the corpus 32. To this end, the index generator 34 may decompose the corpus 32 to create a standard hierarchical topic orientation that is capable of assigning text content to title, headers, topics, subtopics, or any other device that may be employed for  
10 representing a section of text related to a topic, meaning, category, or some other similar abstraction.

US Patent 5,963,203 entitled "*Automatic index creation for a word processor*"  
15 issued to Sotomayor, Bernard, describes methods that may be employed by the index generator 34. For example, Sotomayor describes methods that enable scanning one or more documents to automatically identify key topics and phrases in a document's text, as well as methods to generate an index to those key topics. Similarly, US Patent 5,819,258 entitled "Method and apparatus for automatically generating hierarchical  
20 categories from large document collections", by Vaithyanathan, Shivakumar, Travis, Robert, and Prakash, Mayank, further describes techniques that may be employed by the index generator 34 for determining an index for a corpus. Other techniques known in the art may also be employed by the index generator 34 without departing from the invention.

25 In an alternative practice, the index generator 34 allows an operator to identify the type of corpus 32 being input into the index generator 34. For example, the index generator 34 may present an interface to the operator that allows the operator to identify whether the corpus being presented comprises an encyclopedia, a dictionary, a  
30 textbook, or another known type of reference document. Additionally, the index generator 34 may allow the operator to identify whether the corpus 32 includes a table of contents, an index, chapter heading, or any other representation of the different topics contained within the corpus. In this embodiment, the user may identify, for example, that the corpus 32 comprises an encyclopedia and that the encyclopedia includes a table  
35 of contents that is representative of the index of orthogonal topics maintained within the

5 encyclopedia. In this embodiment, the index generator 34 may process the presented corpus 32 to identify the table of contents for the encyclopedia. This table of contents, in one embodiment, may be formatted into an HTML document that presents the table of contents in an organized format that emphasizes the topics, subtopics and other hierarchical structure of the table of contents. In one process the index generator 34  
10 processes the notation for the table of contents, such as the topic numbering employed, to identify which topics are understood as parent topics and which are understood as main topics and which are understood as subtopics. In a further optional embodiment, the index generator 34 may present the generated index with the orthogonal corpus 40 to the operator to allow the operator to edit or amend the generated index for the  
15 orthogonal corpus 40.

As shown in Figure 3B once the index generator 34 has processed the corpus 32 the index generator 34 may present the index 40 for the corpus 32 to the keyword generator 34. As depicted in Figure 2A, the index 40 may comprise a hierarchical  
20 representation of the orthogonal topics maintained within the corpus 32. This hierarchical representation may include primary topics, such as the depicted topic 38 and a plurality of subtopics 42 that are associated with the primary topic 38.

The keyword generator 48 in one embodiment operates to identify sections of text  
25 of the corpus 32 to be associated with the different topics and subtopics of the index 40. Continuing with the above example, in those practices where the index 40 is generated from the table of contents for the corpus 32, the keyword generator 48 may identify those pages that contain information associated with a topic presented within the index 40. For example, the keyword generator 48 may process the table of content for the  
30 corpus 32 to identify a page number associated with a topic, such as the topic 40 and may analyze the page associated with topic 40 to identify that portion of the page that may be associated with the topic 40. In one embodiment, where headings are presented within the corpus 32, the keyword generator 48 may analyze the page associated with the topic 40 to identify a heading that is representative of the beginning of the  
35 presentation within the corpus 32 of information that is associated with topic 40. For

5 example, the keyword generator 48 may identify a section of text within the associated page that contains the information associated with topic 40 and that is presented in a type font and font size that is representative of a heading. In a subsequent step the keyword generator 48 may identify the location of the heading for the subsequent topic 44 that indicates the beginning of content related to the new topic. The keyword  
10 generator 48 may identify the content that is delimited by the heading 40 and 44 and associate that content as content related to the topic 40.

Once the portion of the corpus 32 that is to be associated with the topic 40 is identified, the keyword generator 48 may process this assigned portion of text to  
15 generate a plurality of search keys, each of which may be representative of a search string for selecting information from a database.

In one embodiment, the system 10 employs the orthogonal construction of the corpus for algorithmic identification of keywords in each topic that distinguish that topic  
20 from its sibling, cousin, ancestor, or descendent topics. Accordingly, the systems described herein may create a set of keywords for a topic that identifies a document associated with a topic and that are keywords which may act to distinguish documents associated with one topic, from documents associated with another topic. For example, the system 10 may employ processes that identify keywords that are associated strongly  
25 with a particular topic. Techniques for creating keywords will be understood from Deerwester, S., Dumais, S.T., Landauer, T.K., Furnas, G.W. and Harshman, R.A. (1990), "Indexing by latent semantic analysis." *Journal of the Society for Information Science*, 41(6), 391-407. Additionally, the system 10 may identify other keywords that act to disassociate a document from one or more other topics. These keywords may be  
30 employed by the system 10 to numerically score over an underlying pool of documents.

The system 10 may employ scoring methods that may utilize traditional information retrieval techniques including the use of synonyms, stemming, frequency, proximity, stop words, hyponyms, and synonyms. If, as in most large document  
35 collections, it is not practical for all documents to be individually scored against the



5 keywords, then a subset of search words is selected to identify candidate documents for  
scoring. Keyword and search terms are identified based on a numerical method that  
apportions words among topics. The goal is for the keywords and search terms to  
10 identify individual blocks of text as found at the nodes of the orthogonal corpus topic  
hierarchy. In an ideal sense, the keywords would be partitioned across the hierarchical  
tree nodes, with each word occurring in only one corpus topic. In addition to word rarity  
among corpus topics, rarity in the underlying document pool may contribute to a word  
being identified as a keyword or search term for a given topic. For example, a keyword  
15 occurring in only one node and only once on the Web, would be a top candidate as a  
keyword and search term.

20 The keyword generator 48 may present as an output, a set of keywords 52 each  
of which may be associated with a topic or subtopic in the index 40. As described  
above, these keywords may be employed to act to distinguish documents associated with  
one topic from documents associated with another topic. Accordingly, as depicted in  
Figure 2B the search keys 52 and associated topics on the index 40 may be presented to  
the search engine 54 for retrieving information from a database or databases of content  
58. To this end, the process 30 applies the search keys to the database 58 to retrieve  
information from the database 58. In one practice as will be described in more detail  
hereinafter, an optional step in process 30 is performed wherein the search keys 52 are  
25 processed to identify a subset of search keys that may be employed for generating search  
queries to one or more search engines, such as Internet search engines, to discover a set  
of documents which are relevant to the topic of interest. Each of the resulting  
documents may be examined in a subsequent step to determine the relevance of the  
content contained within the index. The relevance may be scored, as further described  
30 below, for identifying the relevance of that document, and the score may be employed  
for ordering the sequence in which content is listed as being relevant to a particular  
document.

35 Once the discovered documents are scored for relevance, the process 30 may  
associate portions of the discovered documents to associated topics within the index 38.

5 In a practice wherein the database 58 includes links to URL's for websites, the process  
30 may create a web database that contains website information such as URL's, types,  
dates, topics, contents, size and editor notes that are inserted or updated in the database  
from time to time. Information about the corpus 32 that has been processed, such as the  
10 publisher, the ISBN, and other types of information needed to purchase the book through  
an online transaction may also be stored. The search engine may then provide a  
navigation tool that comprises the HTML representation of the index 38 wherein topics  
and subtopics within the index 38 link to URL's of web content identified as being  
related to the topic or subtopic selected by the user. Optionally, in certain embodiments,  
15 the topics and subtopics may also include links to portions of the corpus 32 that are  
related to the topic selected by the user. In this way, a user may select a topic presented  
by the corpus 32 in view of the information presented by the corpus 32 and related  
information stored on the World Wide Web. In other embodiments, other techniques are  
employed for semantic processing and for determining a topic that may be associated  
with a portion of text within the corpus.

20 The data flow diagram depicted in Figs. 2A and 2B may be implemented in a  
data processing process wherein a data processing program processes the corpus and  
generates an index that links topics in the corpus to information from a data sources,  
such as the Internet. Turning now to Fig. 3, a flow chart illustration of one such process  
25 is depicted. Specifically, Fig. 3 depicts a process 70 for extending a corpus by  
identifying topics covered by that corpus and employing information stored in the  
corpus and related to the topics to identify information in a database that is also related  
to the topic. The process 70 also generates an optional graphical user interface, such as  
the interface depicted in Fig. 1, that includes links for topics listed in the index, and that  
30 may be employed by a user to access the information associated with the listed topics.

35 The process 70 begins with the act 72 of identifying a corpus that is to be  
extended, such as by selecting a publication that contains reference material. In step 72,  
the process 70 transforms, or casts, the corpus into a normal form for processing. In one  
practice, this involves decomposing the document format of the corpus into a standard

5 hierarchical topic orientation with a mechanism for assigning text content to title, headers, topics, and sub-topics. Optionally, stop words, such as the common words "and", "them", and "within", are identified and removed during normalization.

10 After normalization, the process 70 proceeds to step 78 wherein the corpus is processed to identify which portions of the corpus relate to which topic. In one practice wherein the corpus includes a table of contents, the process 70 analyzes the document format of the corpus to locate within the text headings associated with the different topics. For example, as described in the above cited publication US Patent 5,963,203 entitled "*Automatic index creation for a word processor*", header information set off by  
15 HTML tags may be identified to find indicia of topic entries in the document being processed. However, any technique for processing a document to identify the sections of text related to a topic may be applied, including other techniques for analyzing the mark up form language of the document.

20 Proceeding to step 80, the process 70 analyzes the topics to identify a signature that may be understood as representative of the semantics of the topic. In one practice, the process 70 creates a word map per topic and subtopic. To this end the process 70 in step 80 may create a summary representation of the words in the text based on the number of, location of, and proximity of words within each topic and sub-topic. Other  
25 factors may be employed, or substituted for these. Statistics are maintained on different parts of the document structure such as titles, headings, paragraphs, sentences, and image.

30 Table I depicts that several topics may be identified within the corpus. For example, Table I depicts that the processed corpus includes the topics Archaeology, Argentina, Arithmetic, Art, and Astronomy.

<u>Topic</u>
Archaeology

**Argentina****Arithmetic****Art****Astronomy****TABLE I**

The process 70 in one practice may then determine for a given topic, the word count for the words that appear within the portion of text, or other content, associated with the respective topic. This is depicted in Table II, that shows an example of the word count, with stop words removed, for words that appear in the portion of the corpus related to the topic "Astronomy."

**Word counts in Topic Astronomy**

<b>Word</b>	<b>Count</b>
<b>actual</b>	<b>1</b>
<b>ad</b>	<b>1</b>
<b>adopted</b>	<b>1</b>
<b>advances</b>	<b>2</b>
<b>ancient</b>	<b>3</b>
<b>application</b>	<b>2</b>
<b>assigning</b>	<b>1</b>
<b>astronomer</b>	<b>2</b>
<b>astronomers</b>	<b>2</b>
<b>astronomical</b>	<b>5</b>
<b>Astronomy</b>	<b>12</b>
<b>astronomy</b>	<b>11</b>
<b>Astrophysicists</b>	<b>1</b>
<b>astrophysics</b>	<b>1</b>

**TABLE II**

In process 70, after the word count, and other statistics are determined, signatures are generated using orthogonalization. For example, in one practice, given the word counts or word maps for all or a selected subset of topics simultaneously, the process 70

5 assigns a weight based on word count to each word within each topic or subtopic. Where using word counts the weight may be defined as the count. When using the word map, the weight of a word in a topic or subtopic may be assigned by an intra-document scoring function. Any suitable technique may be employed for performing intra-document scoring. These signatures may be edited or cleaned manually to enhance the topical relevance and precision of the subsequent search and scoring process. Table III depicts an example signature for the topic "Astronomy."

Signature for topic Astronomy	
Word	Count
Astronomy	23
earth	9
bodies	5
Astronomical	5
universe	4
celestial	3
circle	3
Observational	3
sky	3
Stars	3
Ancient	3
Daily	3
Heavenly	2
Astronomers	2
planet	2
relative	2
moved	2
heavens	2

TABLE III

After determining a signature, the process 70 may perform the optional step, step 82, of applying synonym Groups. In this optional step, the process 70 extends the

signatures with synonym groups. To this end, words are replaced by groups of word substitutes having similar or identical meaning. Table IV depicts such an extension.

Astronomy Signature Post Synonym Reduction	
Word	Count
Astronom	30
earth	9
bodies	5
universe	4
celestial	3
circle	3
Observational	3
sky	3
stars	3
ancient	3
daily	3
Heavenly	2
planet	2
relative	2
moved	2
heavens	2

TABLE IV

After step 82, the process 70 may proceed to step 84, wherein the process reduces the signature to Keyword sets, optionally tailored for the search. The set of documents to be scored against a topic is preferably identified and manageable in size. The web for example is a large a document set to collect up and score against all web documents. Accordingly, in one practice traditional large scale search engines, such a Lycos and Alta Vista, may be used to identify a set of candidate relevant documents using a keyword set for search. Which subset of the Signature and synonym groups is included in the Keyword set may be determined based on a variety of measures

5 including corpus document word count of the word and general frequency of the word.  
An example is presented in TABLE V.

Word	Count
Astronomy	30
Earth	9
Bodies	5
Universe	4
Celestial	3
Circle	3
Observational	3
Sky	3
Stars	3
Ancient	3
Daily	3

20 **TABLE V**

25 The keyword set may be applied to a search mechanism to pull in multiple  
discovered documents based on the keyword set. This may occur in step 88. For  
example, the query Find: Astronomy or Astronomical or Astronomers or earth or bodies  
or universe or celestial or circle or Observational or sky or stars or ancient or daily; may  
be generated from the keyword set and applied to the search mechanism to discover  
documents related to the selected topic.

30 After the step 88, the process 70 may proceed to step 90 for scoring of the  
discovered documents. The many discovered documents returned from a search  
function may be assigned individual scores against the corresponding corpus topics and  
subtopics. Scoring may be based on multiple tunable metrics and rules including  
35 functions over the word count or word\_map data structures. The score of topical overlap  
between two documents as a baseline is measured as a dot product of word counts or  
word frequencies within those documents).

5

## Scoring example

10

15

20

Word	Count in Astronomy	Count in discovered document	Score Contribution
Astronom	30	2	0.555556
earth	9	1	0.083333
bodies	5		
universe	4	5	0.185185
celestial	3		
circle	3	2	0.055556
Observational	3		
sky	3		
stars	3		
ancient	3	1	0.027778
daily	3	2	0.055556
Heavenly	2		
planet	2		
relative	2		
moved	2	1	0.018519
heavens	2	1	0.018519

25

30

After step 90, the process 70 proceeds to optional step 92, wherein the topic hierarchy and set of associated documents may be presented directly through an HTML or graphical user interface, such as the interface depicted in Fig. 1. Alternatively, content may be delivered through software API's (application program interfaces) to allow integration of output content with other content. Content may be navigated by walking the directory tree structure, or by keyword searching over the directory structure trees, corpus content, or discovered document content. Search results may point to topic paths or discovered documents,

35

40

Fig. 4 depicts one embodiment of the system 100 according to the invention. Specifically Figure 4 depicts a functional block diagram that shows a system 100 that allows a surfer 102 to access a user interface 104 that couples to a database system 108. The database system 108 further couples to an OCI processor 112 that accesses a database of corpora 114 and a plurality of search engines 118. The database system 108 further couples to an application programming interface access layer 120 and through the API 120 can access a portal/search client 122. Additionally, the API 120 may also couple to a scoring mechanism 124.



5           More particularly, Figure 4 depicts that a user 102 such as an Internet user may access a user interface 104, that may be similar to the user interface depicted in Figure 1. As shown in Figure 1 the user interface 10 may present to the user 102 a list of topics 112. The user 102 may select a topic from the index 112. As described with reference to Figure 1, the selection of a link directs the user interface 104 to retrieve information from the database system 108. The database system 108 processes the users request from user 102 for information related to the selected topic.

10           The database system 108 may be any suitable database system, including the commercially available Microsoft Access database, and can be a local or distributed database system. The design and development of suitable database systems are described in McGovern et al., *A Guide To Sybase and SQL Server*, Addison-Wesley (1993). The database 108 can be supported by any suitable persistent data memory, such as a hard disk drive, RAID system, tape drive system, floppy diskette, or any other suitable system.

15           As further depicted by Figure 4 that database system 108 may communicate with the OCI mechanism 112. The OCI mechanism 112 may be, in one embodiment, a computer process capable of implementing a process such as process 70 depicted in Figure 3. The OCI mechanism can be realized as a software component operating on a conventional data processing system such as a Unix workstation. In that embodiment, the OCI mechanism can be implemented as a C language computer program, or a computer program written in any high level language including C++, Fortran, Java or basic. Techniques for high level programming are known, and set forth in, for example, Stephen G. Kochan, *Programming in C*, Hayden Publishing (1983). Accordingly, the OCI mechanism 112 may be employed by a system administrator to process corpora stored within the database 114. As discussed with reference to Figure 3, the processed corpora results in a graphical user interface that may be stored within the database mechanism 108 and accessed by the user 102 through the topic navigator 104. Additionally, the OCI mechanism 112 may generate for the processed corpora of database 114 a set of links or pointers to content that corresponds with different listed

5 topics within the index of the processed corpora. The OCI mechanism 112 may also store these associated links within the database system 108.

To this end, the OCI mechanism 112 may couple to one or more search engines 118 that allow the OCI mechanism 112 to retrieve content from a database source. In the depicted embodiment of Figure 4, the database source that search engines 118 access is the World Wide Web 106. In this embodiment, the user interface 104 also couples to the World Wide Web 106 so that links activated by the user that relate to URL's of content stored on the World Wide Web 106 may be directly accessed by the user 102 through the user interface 104 through the connection between the user interface 104 and the World Wide Web 106. Figure 4 further depicts that the database 108 communicated with an API layer 120. As shown in Figure 4 the API layer sits between the portal search client 122 and the database system 108 and also sits between the scoring mechanism 124 and the database system 108. Accordingly, a portal search client such as the Yahoo site may access the database system 108 through the API layer to provide users with access to an index linked to content on the World Wide Web.

Similarly, Figure 4 depicts the scoring mechanism 124. The scoring mechanism 124 may be a computer process that accesses the database system 108 through the API 120. The scoring mechanism may perform data mining for identifying topics that are to be associated with different websites. In this way, the database system 108 may be employed for categorizing web sites according to their content. Thus, the system 100 depicted in Figure 4 provides a system for categorizing information stored on the World Wide Web. The system described in Fig. 4 may operate on any suitable computer hardware, such as PC compatible computer systems, Sun workstations, or any other suitable hardware. The list of topics and the associated documents, or links to documents may then be stored in a relational database, or any suitable database with proper indexing for allowing rapid accessing of the data stored therein.

5           Once the system is operating, the system may be employed to provide a set of tools, such as that may operate as stand alone applications for single users, or that may be tools provided as client/server programs over a network. The tools may be provided as a collection of functions incorporated into an integrated research tool, or may co-exist as individual functions in a separate application.

10           Further embodiment, the systems and methods described herein may be employed for organizing a plurality of corpora into an indexed format that may be presented as a graphical user interface for a user to allow a user to access information related to the contents of a plurality of corpora. For example, figures 5 through 9 depict the operation of a system that processes a plurality of text, such as reference texts. Accordingly, the system may be employed for the automatic creation of a topically organized book catalog, such as a catalog of reference books, with navigation, search, click through to external documents such as web documents, with information purchasing interfaces also. For example, figure 5 depicts a graphical user interface that presents to a user a plurality of topics each having a set of books within the topic. For example, the figure 5 depicts a topic reference that includes a set of encyclopedias and dictionaries within that reference. By activating the reference link, the user may be presented with the user interface shown in figure 6. Figure 6, the individual references presented under the reference topic of figure 5 are outlined for the user allowing the user to select what type of reference the user would like to view.

25           For example, the user may select from encyclopedias, dictionaries, academic and learned society publications and other such publications. After making a selection figure 6, the user may be presented with the different books under each category. The example presented in figures 5 through 9 shows that upon activating the link for encyclopedias, the user is presented with the different encyclopedias that have been processed by the system according to the invention. Upon selecting a link, such as the link for the Encyclopedia Britannica, the user may be presented with the interface shown in figure 8 that lists the different topics covered by the Encyclopedia Britannica. At this level, the process now proceeds as described above, with reference to figures 1 through 4

5 wherein the individual topics maintained within the Encyclopedia Britannica may be employed for accessing contact, such as web contact particularly associated with the individual topics.

10 Those skilled in the art will know or be able to ascertain using no more than routine experimentation, many equivalents to the embodiments and practices described herein. For example, the systems and methods described herein may be employed for providing encyclopedia (i.e., corpus) extender. An encyclopedia (as an archetype example of an orthogonal corpus) may be automatically extended by application of the systems and methods described above, to include links into the World Wide Web, or  
15 other database, via searching or meta-searching over the Web. The breadth and depth of the corpus enables a high quality, high coverage database of web links, with the web links organized according to the location in the topic hierarchy whose text was used to generate them. Such links may provide geographical maps, histories of topics of interest, access to theses and other types of information. Other applications include  
20 web book companions wherein the system processes a book, including a fictional work, a non-fictional work, or a reference book, through this system will allow automated construction of topical web sites as Web Companions to individual books. For example, a book such as the Hunt for Red October may be processed by the systems described herein to create links into the Web to documents associated with concepts from the  
25 book, such as links to the Navy Submarine division, links to topographic maps of the ocean floor, links to Russian Naval History, and other similar links.

A search engine extension may be provided by accessing the database 108 through the API. Thus a user may do a search on a web search engine, they may want to  
30 refine their search or get a second search opinion. Given a broad topic database such as that created in the Encyclopedia Extender application described above, refinement of a user's intended topic is enabled – through keyword-based narrowing, web link browsing, and display of proximal or correlated topics in the corpus topic hierarchy. For e-commerce, the systems described herein book/article browser/seller. Browsing over the  
35 topic hierarchy may provide indexes into books or articles for sale.

5 Additional applications can include a user Interface. The user interface allows users to view Web links through the topic hierarchies defined by the corpus. The topic hierarchy on the left lists the topics as per the corpus. The user may select keywords from the corpus outline, or from provided sample text inside the corpus documents, to better focus and score the topic. Users may augment the search terms or keywords with their own keywords or selected synonyms to more specifically tailor a concept to a need. Searching across the corpus or across the referenced links may include synonyms, stemming, frequency, proximity, stop words, hyponyms, and synonyms.

10 Additionally, authoring toolkits may be provided that allow publishers, editors, and authors to create corpus extensions and associated applications. For example, the systems and methods described herein may be employed to create development kits that publishers may use to index a book and create a web site that acts as the book companion described above.

15 It may be noted that human oversight or auditing of the document scoring and database may be done in order to augment the purely automated document selection. This may be done on a sampling basis for quality control. Different levels of sensitivity to content or product price points may be implied by different levels of human quality control. Moreover, it will be noted that the system described above has been done so with reference to documents stored on the Web. However, it will be understood by those of ordinary skill in the art, that the Web is being used here as a metaphor for any electronic document archive, and the systems and methods described herein are not limited to the Web.

20 Accordingly, it will be understood that the invention is not to be limited to the embodiments disclosed herein, that other applications, such as information mining may be practiced with the systems and methods described herein, and that the invention is to be understood by the following claims which are to be interpreted as broadly as allowed under the law.

5

We claim:

10

1. A method for processing a body of reference material to generate a directory for accessing information from a database, comprising

15

processing the body of reference material to identify a hierarchical organization of a plurality of topics,

associating with at least one of said topics a portion of the reference material,

processing said assigned portion of the reference material to generate a plurality of search keys representative of search strings for selecting information from the database,

applying the search keys to the database to retrieve information from the database, and

creating an association between the at least one topic and the information retrieved from the database.

20

2. A method according to claim 1, further comprising

creating a graphical interface representative of the identified hierarchical organization of a plurality of topics for allowing a user to access information retrieved from the database and having an association with a topic.

25

3. A method according to claim 1, wherein processing the body of reference material includes processing a body of reference material selected from the group consisting of an encyclopedia, a dictionary, a text book, a novel, a newspaper, a web site, and [ Henry what else are we interested in processing?].

30

4. A method according to claim 1, wherein processing the body of reference material includes processing the body of reference material to identify a hierarchical organization of a plurality of substantially orthogonal topics.

35

5. A method according to claim 4, wherein processing the body of reference

- 5 material includes identifying a table of contents for the body of reference material.
6. A method according to claim 4, wherein processing the body of reference material includes identifying chapter headings.
- 10 7. A method according to claim 4, wherein processing the body of reference material includes identifying definition entries in a dictionary.
8. A method according to claim 4, wherein processing the body of reference material includes normalizing the identified hierarchical organization of a plurality of topics. [ Henry - here we are interested in any things we do to process the hierarchical organization to make the topics more orthogonal].
- 15 9. A method according to claim 1, wherein processing the assigned text includes generating a word map representative of a statistical analysis of words contained in the assigned text.
- 20 10. A method according to claim 9, wherein generating the word map includes performing a word count process for determining word frequency of a word within the assigned text, and for employing the word frequency for determining relevance of a word to the associated topic.
- 25 11. A method according to claim 1, wherein processing the assigned text includes processing the assigned text to identify a set of keywords having an associated measure of intra document orthogonality.
- 30 12. A method according to claim 1, wherein processing the assigned text includes identifying a set of synonyms for extending the search keys.
- 35 13. A method according to claim 1, wherein processing the assigned text includes

5                    selecting a subset of the search keys having a predetermined measure of correlation to the topic.

10                   14.    A method according to claim 1, wherein applying the search keys to the database includes applying the search keys to an Internet search engine to discover documents related to the search keys.

15                   15.    A method according to claim 1, wherein applying the search keys to the database includes applying the search keys to a meta-search engine to discover documents related to the search keys.

                    16.    A method according to claims 14 or 15, wherein applying the search keys to the database includes processing text of the discovered documents to identify documents related to the topics associated with the search keys.

20                   17.    A method according to claim 1, wherein creating an association between the at least one topic and the information retrieved from the database includes capturing a location pointer associated with the information retrieved from the database.

25                   18.    A method according to claim 1, wherein creating an association between the at least one topic and the information retrieved from the database includes creating a data structure for the topic for storing location pointers associated with information retrieved from the database.

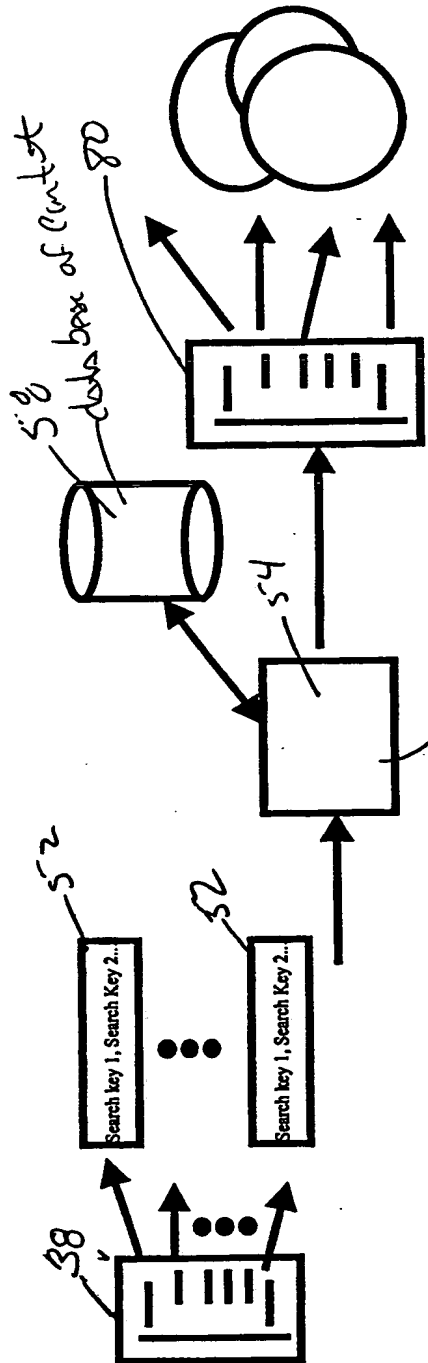
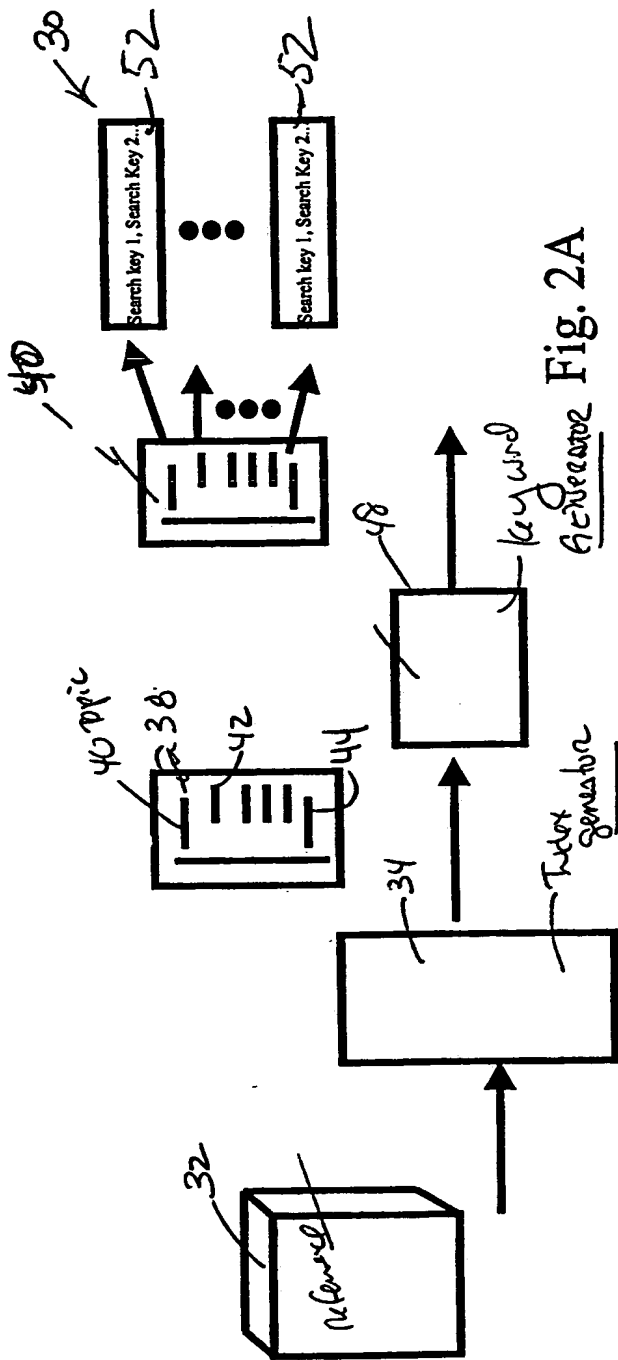
30                   19. A system for organizing a collection of documents, comprising  
                    an orthogonal corpus of information arranged according to an index of topics,  
                    a keyword generator for generating a set of keywords representative of a document's association with a topic in the index of topics,  
                    a scoring system for processing documents within the collection of documents, to  
35                   associate with at least a portion of the documents, a score representative of the document's association to a particular topic, and



5                   a graphical representation for depicting at least a portion of the index of topic and  
having respective portions of the graphical representation linked to documents  
associated with a respective portion of the index of topics.

20. A system for extending the content of a first document, comprising  
10                   a parser for selecting terms with said first document to be extended,  
                  an orthogonal corpus of information arranged according to the index of topics,  
                  a keyword generator for generating a set of keywords representative of a  
document's association with a topic in the index of topics, and  
                  a linking system for processing documents within a collection of documents, to  
15                   associate with at least a portion of the documents, a score representative of the  
document's association to a particular topic, and for providing said first document with  
links into said collection of documents for extending the content of said document.





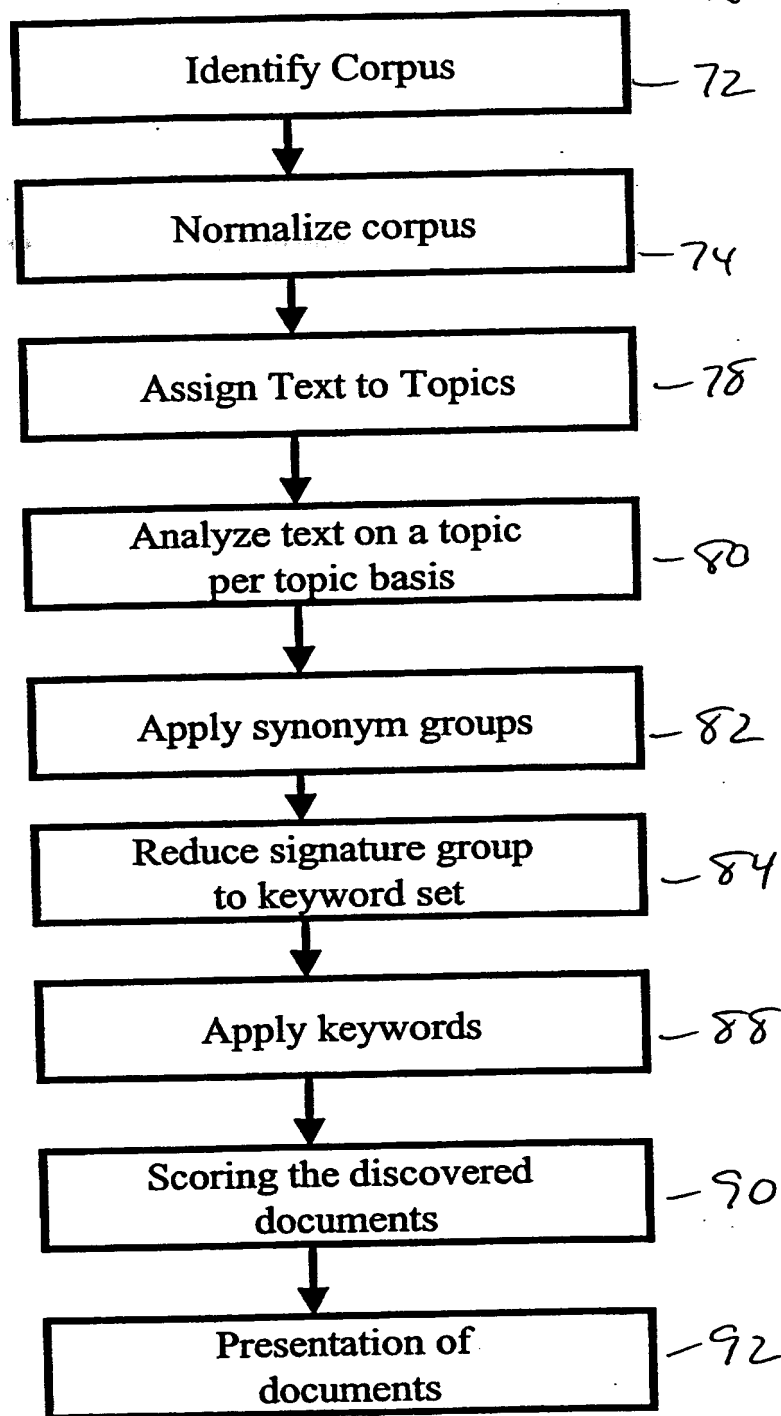


Fig. 3

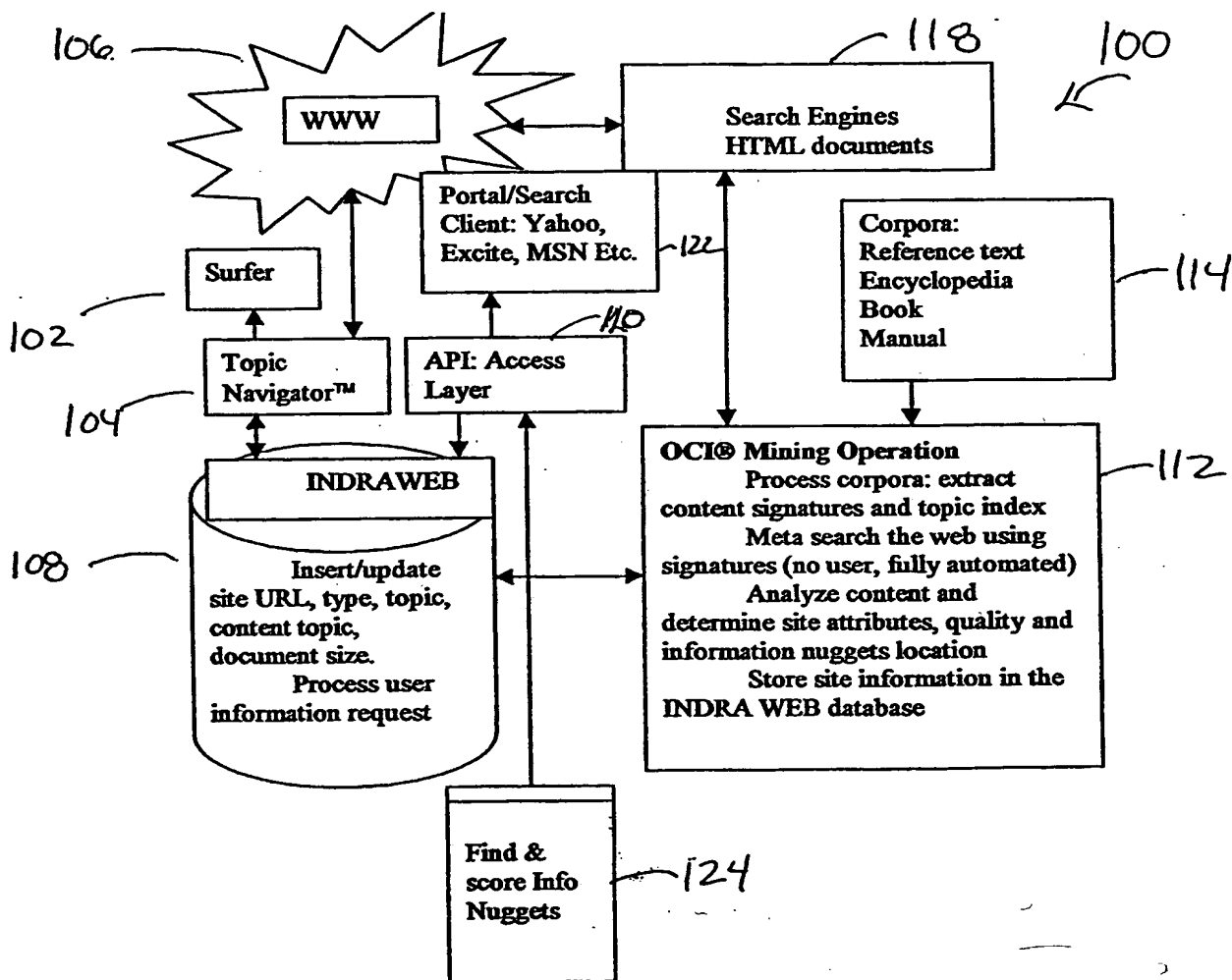
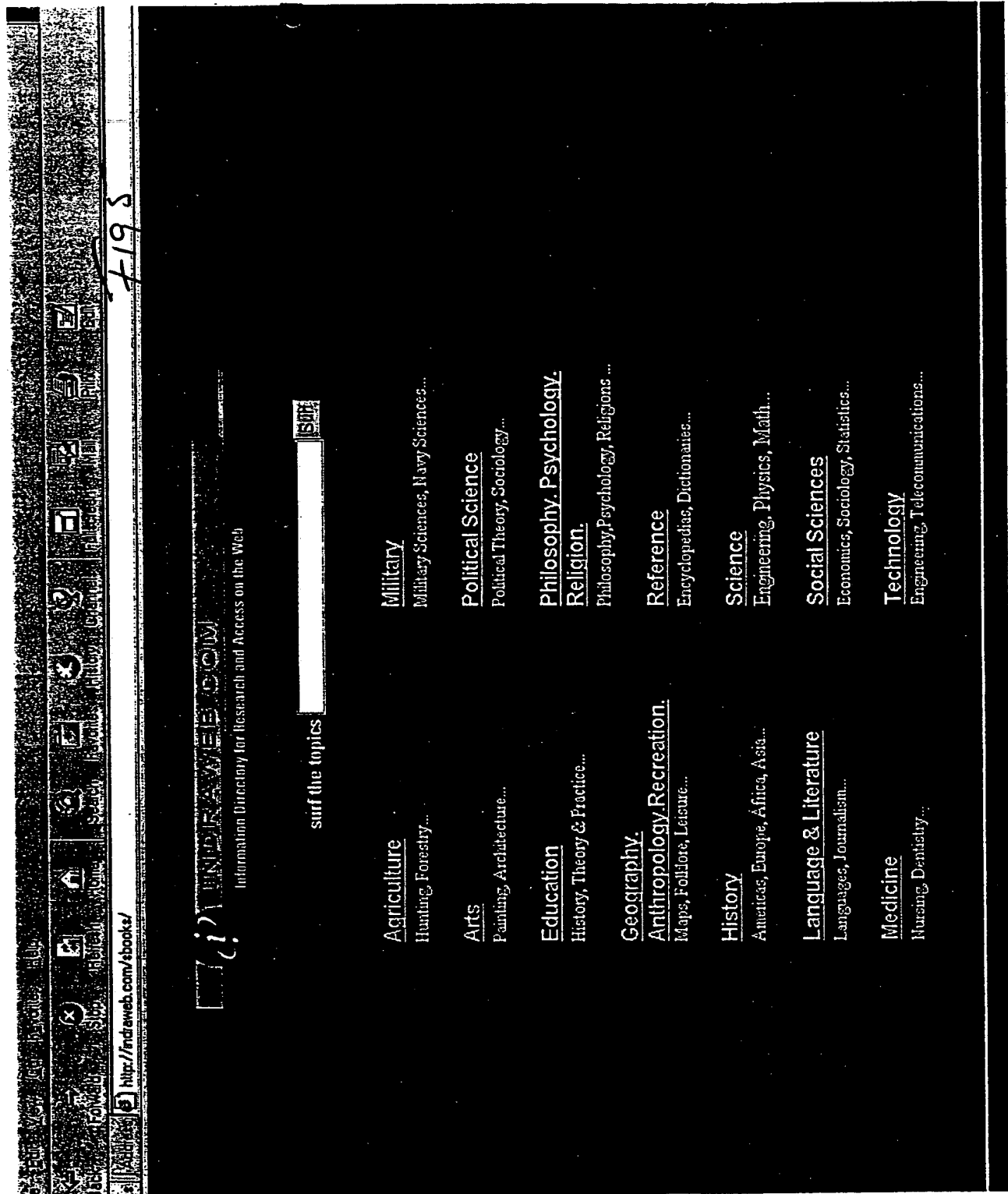
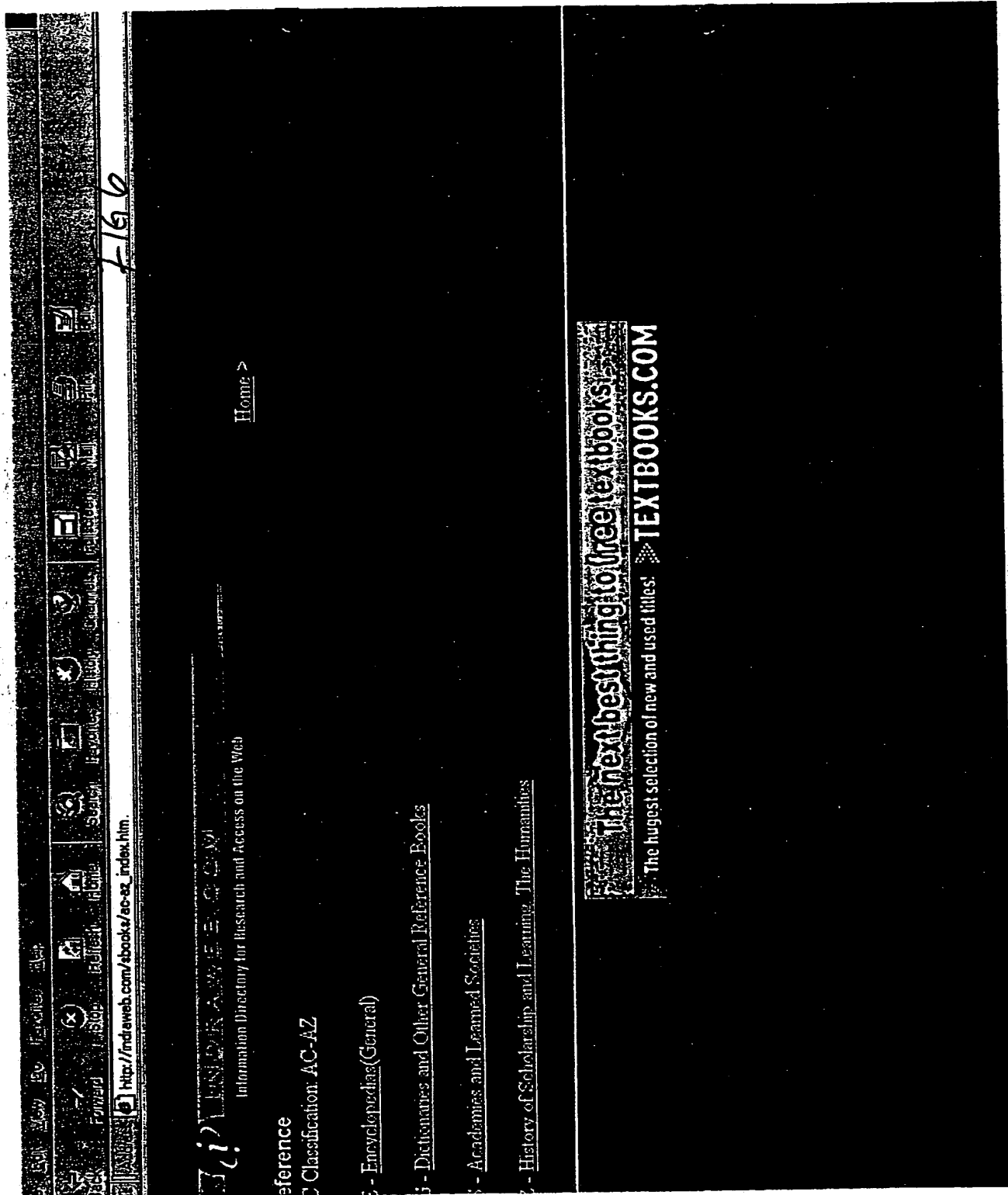
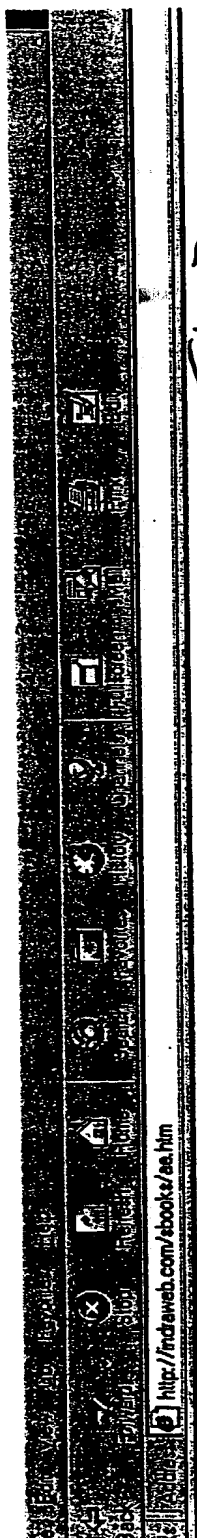


Fig. 4







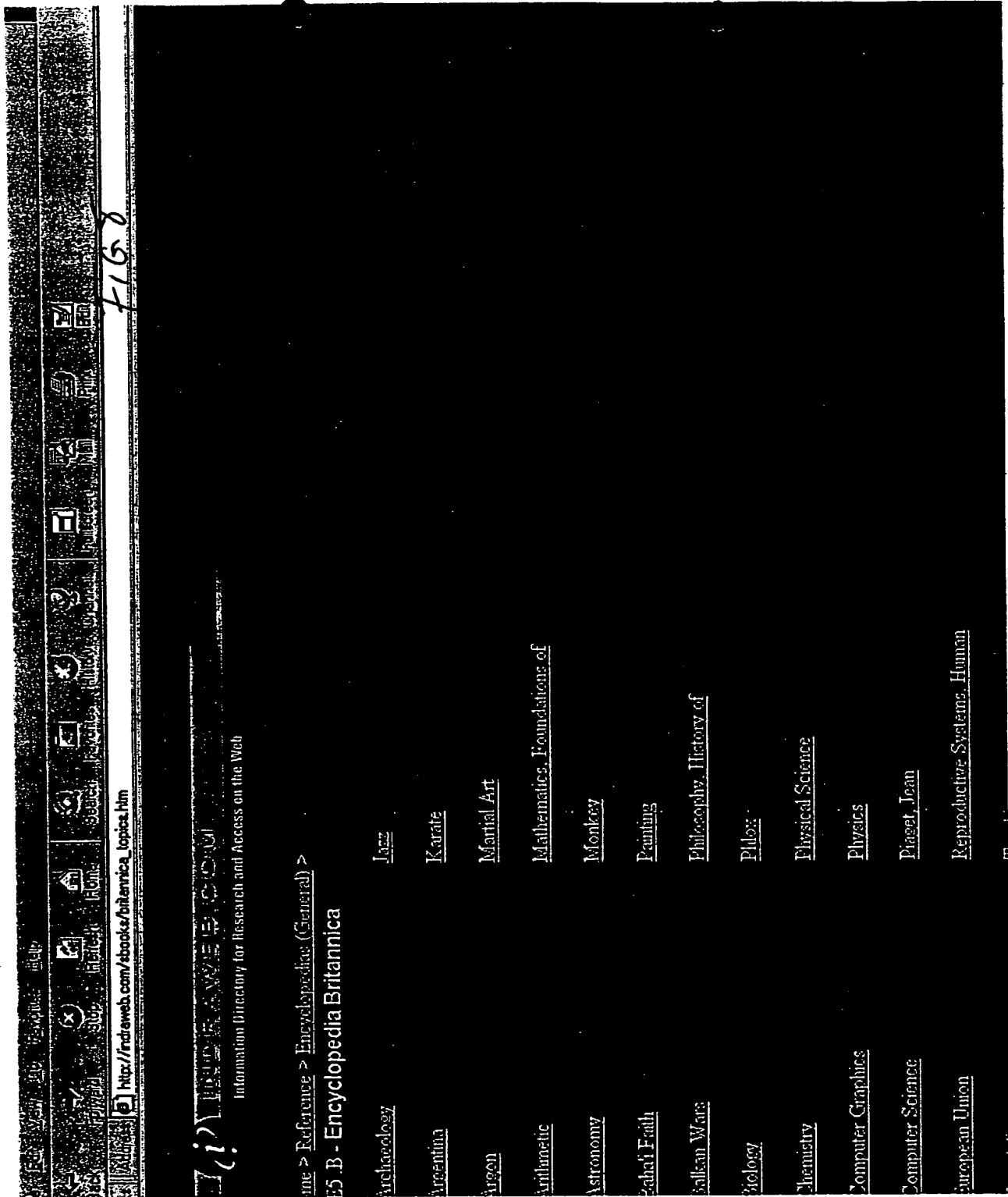
Home > Reference >

E - Encyclopedias (General)

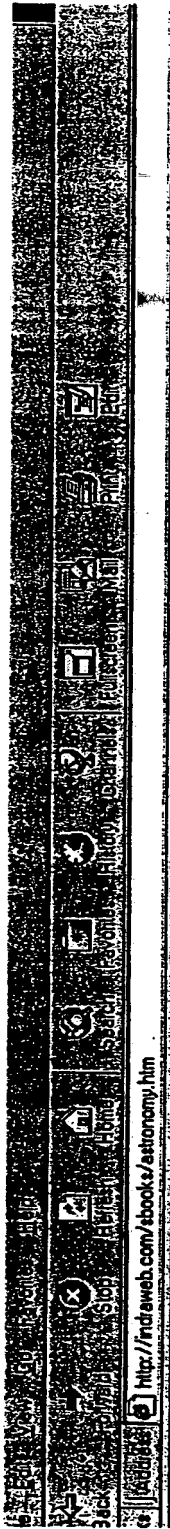
name bonum : a fourteenth-century encyclopedia of universal knowledge : British Library MSS Royal 6 E VI-E VII	1996	AE2 L463 S26	H. Miller Publishers
cademic American encyclopedia.	1998	AE5 A23	Grolier Inc.
ncyclopedia Britannica	1997	AE5 B	Routledge
ne Cambridge paperback encyclopedia	1995	AE5 C36	Cambridge University Press
ambridge paperback encyclopedia	1993	AE5 C36	Cambridge University Press
ollier's encyclopedia : with bibliography and index	1997	AE5 C683	Collier's
ne Encyclopedia Americana	1999	AE5 E333	Grolier Incorporated
ncyclopedia of knowledge	1991	AE5 E478	Grolier
amily encyclopedia	1997	AE5 F36	Oxford University Press
unk & Wagnalls new encyclopedia	1996	AE5 F83	Funk & Wagnalls
ne Macmillan compact encyclopedia	1991	AE5 M24	Macmillan
lerit students encyclopedia	1993	AE5 M38	Macmillan Educational Co.
ew standard encyclopedia	1998	AE5 N64	Ferguson Pub. Co.
xford illustrated encyclopedia	1993	AE5 O94	Oxford University Press
ebster's family encyclopedia	1992	AE5 W35	Meriam-Webster, Inc.
ne World Book encyclopedia	1999	AE5 W55	World Book
ne World Book encyclopedia of people and places.	1998	AE5 W563	World Book

F167





9/9



me > Reference > Encyclopedias (General) > Encyclopaedia Britannica >

## 55 B Encyclopaedia Britannica: Astronomy

astronomy earth bodies astronomical universe celestial circle observational sky stars ancient dally heavenly astronomers planet  
 search: relative moved heavens ray ptolemaic copernicus revolution recognized electronic advances position rad (Any word)

und: 277 document(s) on The Web

Sort: Score

### Imagine the Universal Dictionary

Please allow the whole page to load before you start searching for an entry. Otherwise, errors will occur. [A B C D E F G H I J K L M N O P Q R S T U V W X Y Z]

Found by: Infoseek, WebCrawler

<http://imadline.arts.utoronto.ca/dictionary.html>

72%

### APOD Glossary

Astronomy Picture of the Day Glossary Cepheid: A pulsating variable star. This type of star undergoes a rhythmic pulsation as indicated by its regular pattern of changing brightness as a function of time.

Found by: WebCrawler

<http://anbspod.astro.nyu.edu/~gcn/glossary.html>

70%

### ASTRONOMY

In this page I've written something about. ASTRONOMY. Version en Castellano. WE CAN ALWAYS LEARN SOMETHING NEW I Astronomy is the science of heavenly...

Found by: AltaVista

<http://www.gccolleges.com/Yosemite/Trailer7812/astronomia.html>

<http://chandra.bgsu.edu/~gcd/Spacetime6.html>

Back to the course home page. Back to Table of Contents. Chapter 6 The Transition from Aristotle to Newton I. Introduction The purpose of this chapter is to describe, in broad outlines, the contributions to the theory of space and time which were...

Found by: HotBot

<http://chandra.bgsu.edu/~gcd/Spacetime6.html>

70%

### Expanding Universe: Astronomy

520 : 525 - Astronomy and Allied Sciences. 520 522 523 525. General astronomy sites, periodicals, associations, planetariums Observatories...

Found by: AltaVista

<http://www.mtl.utoronto.on.ca/courses/bcd/astronomy/520-25.htm>

70%

### The Atheism of Astronomy

I tharv: Historical Documents: Wnnkeav Tallar. Atheism Of Astronomy The Atheism Of Astronomy: A Refutation Of The Theory That The Universe Is Governed By

70%

CORRECTED VERSION

(19) World Intellectual Property Organization  
International Bureau(43) International Publication Date  
19 October 2000 (19.10.2000)

PCT

(10) International Publication Number  
WO 00/62198 A2(51) International Patent Classification<sup>7</sup>: G06F 17/30

(21) International Application Number: PCT/US00/10003

(22) International Filing Date: 13 April 2000 (13.04.2000)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:  
60/129,103 13 April 1999 (13.04.1999) US

(71) Applicant: INDRAWEB.COM, INC. [US/US]; P.O. Box 323, Concord, MA 01742 (US).

(72) Inventors: KON, Henry: 91 Bow Street, Lexington, MA 02420 (US). BURCH, George: P.O. Box 323, Concord, MA 01742 (US).

(74) Agents: KELLY, Edward, J. et al.; Foley, Hoag &amp; Eliot, LLP, One Post Office Square, Boston, MA 02109 (US).

(81) Designated States (*national*): AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CR, CU, CZ, DE, DK,

DM, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZW.

(84) Designated States (*regional*): ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

## Published:

— without international search report and to be republished upon receipt of that report

(48) Date of publication of this corrected version:

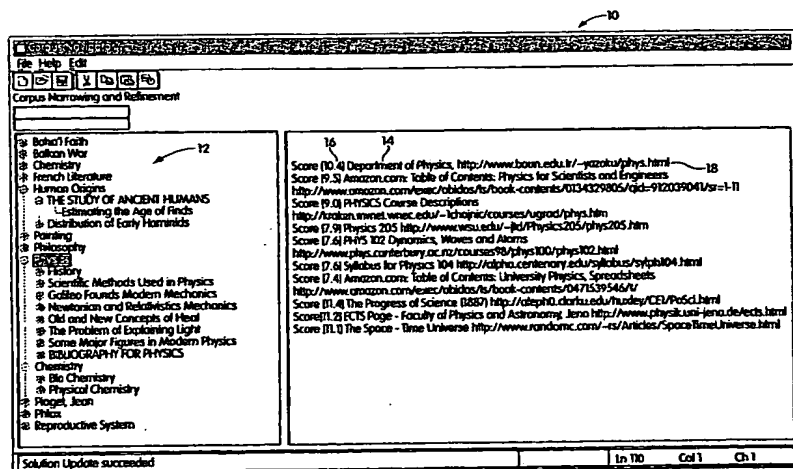
29 November 2001

(15) Information about Correction:

see PCT Gazette No. 48/2001 of 29 November 2001, Section II

[Continued on next page]

(54) Title: SYSTEMS AND METHODS FOR EMPLOYING AN ORTHOGONAL CORPUS FOR DOCUMENT INDEXING



(57) Abstract: The invention provides for indexing and cataloging of content on the Internet, as well as from other stores of information, may be performed by applying a process that employs an orthogonal corpus, or corpora, of information, such as an Encyclopedia. To this end, the processes described herein identify the topics discussed within the corpus. The process also identifies within the corpus a set of keywords that are relevant to the topics presented in the corpus. The keywords associated with a topic may be employed to identify documents stored in another database that are related to the topic. A graphical representation of the index of topics found in the corpus may then be generated, with individual topics operating as links to these related documents. Thus, a user interested in reviewing content in the corpus related to a certain topic, may also activate a link in the graphical representation of the index to access other documents that have been identified as related to the topic of interest to the user.

WO 00/62198 A2

*For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.*

# SYSTEMS AND METHODS FOR EMPLOYING AN ORTHOGONAL CORPUS FOR DOCUMENT INDEXING

## Field of the Invention

The invention provides systems and methods for organizing a collection of information, such as a database of documents, and more particularly, to systems and methods for classifying information with respect to a set of independent concepts, and applications of the classified information.

## Background of the Invention

The computer-based storage of information has become the norm for our knowledge-based society, and today huge databases of computer-stored information are available to researchers. For example, through the World Wide Web service, the Internet has become an enormous repository of readily accessible information. However, the sheer volume of information maintained on the World Wide Web becomes problematic when someone wishes to identify those resources, or documents that are maintained on the Internet and that are related to a certain topic. The problem is that a simple search request for information stored on the Internet and related to a certain topic returns an enormous volume of information, most of which is extraneous and irrelevant. This problem arises in part from the difficulty that existing Internet retrieval systems have with prioritizing and categorizing documents stored on the World Wide Web. Today, most systems for organizing content on the web work by searching through a document stored on the web, and associating with the document a set of key words that represent the information contained on the reviewed pages. Accordingly, documents can be organized according to the keywords that appear in the document.

5                   However, lists of keyword, although helpful, lack an identifiable hierarchy of  
concept and therefore fail to provide a topic-based system for organization and  
document retrieval. Thus, keyword lists are less powerful tools for document research  
than topic-based hierarchal systems of document organization, such as tables of content,  
or topic indexes. Today there is a need for a "table of content" for the World Wide Web.  
10                  Such a table of content would provide a more powerful metaphor for using the Web as a  
research tool and would provide a flexible platform on which other research tools could  
be built. As such, there is a need for improved systems for organizing information  
stored in a computer database, such as the World Wide Web.

### 15                  Summary of the Invention

                  The systems and methods described herein provide systems for document  
indexing and scoring of content on a computer database, such as the World Wide Web.  
The systems generally include an orthogonal corpus that may comprise a collection of  
blocks of text, and that may be employed to index and score textual information for  
20                applications in retrieving, classifying, or browsing over a set of documents.

                  An orthogonal corpus, as the term is employed herein, may be understood to  
encompass, without being limited to, any collection of blocks of text that are outlined or  
referenced by a table of contents, topic index, chapter heading or other topical indicia  
where each topic either stands alone or is an identified subpart (subtopic) of another  
25                topic, forming a tree of topics and their descendant subtopics. Encyclopedias, text and  
reference books, periodicals, web sites, dictionaries, thesauri, the library of congress, the  
Dewey decimal system, and glossaries are examples of, surrogates for, or extenders of  
orthogonal corpora. A set of topics is understood as orthogonal in the sense that  
substantially every member topic (e.g., chapter or article) covers a different concept or  
30                substantially different concept than any other topic under the same ancestor topic in the  
tree. A topic in one practice may be assumed to include or not include, its ancestor  
topics or underlying subtopics.

                  The orthogonal construction, or decomposition, of a corpus provides for  
algorithmic identification of keywords in each topic, which distinguish it from its  
35                sibling, cousin, ancestor, or descendent topics. Keywords may be employed to

5 numerically score over an underlying pool of documents. Alternatively, if more practical for a given application, such as when working with a large document collection, rather than individually score all documents in the collection against the keywords, then a set of search words may be generated to identify a subset of candidate documents for scoring.

10 Parameters employed during the scoring process may relate to the identification of keywords and their refinement into a set of search words, with weightings for associated document or sub-document scoring. The scoring, search term, and keyword methods may employ conventional information retrieval techniques including the use of synonyms, stemming, frequency, proximity, stop words, hyponyms, and synonyms.

15 For purposes of clarity certain terms will now be described, although the understandings set forth are not to be understood as limiting and are only provided for purposes of achieving clarity by way of providing examples. The term "word" as employed herein may be understood to encompass a lexical type found in a common or specialty dictionary of any language. The term "phrase" as employed herein may be understood to encompass any sequence of one or more words. Heretofore for simplicity we use "Word" to mean "Word or Phrase." The term "synonym group" shall be understood to encompass a set of words which may be used as alternates for a given word. Each word in a synonym group has a similar or identical meaning. The term "topic" shall be understood to encompass textual content typically having a title, having corresponding text, concerning a single topic, or covering a set or tree of related subtopics. The term "subtopic" shall be understood to encompass a block of text within a Topic. Typically, a subtopic may be identified by its subtopic header or other outline indicator. In different calculation contexts, subtopics may or may not be included as part of parent topics. The term "word\_count" maybe understood as an integer count of the number of times a word or a word in its synonym group occurs in a given topic or text area, potentially including text in the title and headers and any text elements in that text. The term "word\_frequency" may be understood to encompass the word count in a text area divided by the number of words total in that text. A word\_map is a representation of textual content within a text area that is more precise than a word\_count. A word\_map may describe a word's relative location in the text, its linguistic type or contexts of use,

5 its prominence indicators such as use in a title or highlighting fonts. The term  
orthogonal corpus may be understood to encompass a collection of topically organized  
information referenced by a table of contents and/or index, where each topic is clearly  
10 identified as a SubTopic of a topic or else stands alone. Nodes in the table of contents  
tree may represent topics. The information may be understood as orthogonal in the  
sense that a stand-alone topic (e.g., chapter or article) covers a substantially different  
concept than any other stand-alone topic, and any subtopic expresses a substantially  
15 different concept from any other subtopic within the same parent topic. The term  
document may be understood to encompass formatted textual content with topic  
beginnings, endings, and marked hierarchy. A document may contain one or more  
topics and may include subtopics. A corpus may include one or more documents. The  
relationship between documents and topics is not mandated, though in some  
embodiments each document to represent one top-level topic along with its subtopics.  
The term "discovered document" may be understood to encompass a document (or a set  
20 of documents such as a web site or portion of a web site) which is being scored. Scoring  
of a discovered document may be relative to one or more corpus documents or corpus  
topics. In one practice, scoring measures the degree of topical relevance to the corpus  
topics. The discovered document will often be a member of a search result set.

25 More particularly, the systems and methods described herein include methods for  
processing a body of reference material to generate a directory for accessing information  
from a database. These methods may comprise processing the body of reference  
material to identify a hierarchical organization of a plurality of topics. Additionally, the  
processes may include the step of associating with at least one of the topics a portion of  
30 the reference material and processing the assigned portion of reference material to  
generate a plurality of search keys representative of search strings for selecting  
information from the database. The process may then apply the search keys to the  
database to retrieve information from the database and may create an association  
between the at least one topic and the information retrieved from the database.

35 In an optional step, the methods described herein may create a graphical interface  
that is representative of the identified hierarchical organization of a plurality of topics



5 for allowing a user to access information retrieved from the database and having an association with the topic. Accordingly, the user may be provided with a graphical interface that allows the user to activate, typically by clicking with a mouse, a graphical representation of a topic to identify a set of links to content, such as web pages that are associated with the topic selected by the user.

10 In one practice, processing the body of reference material includes processing a body of reference material that has been selected from the group consisting of an encyclopedia, a dictionary, a text book, a novel, a newspaper, or a website. Processing the material may include identifying a hierarchical organization of a plurality of substantially orthogonal topics. This may include identifying a table of contents for the  
15 body of reference material, identifying an index for the reference material, identifying chapter or subchapter headings within the reference table, identifying definition entries within a dictionary, and other similar operations that identify different topics that occur within the reference material.

20 Optionally, when processing a body of reference material, the process may normalize the identified hierarchical organization of the plurality of topics.

In one practice when processing the assigned text, the process includes a step of generating a word map that is representative of a statistical analysis of the words contained in the assigned text. Generating the word map may include performing a word count process for determining word frequency of a word within the assigned text  
25 and for employing the word frequency for determining the relevance of a word to the associated topic. Processing the assigned text for different topics may also include a step of identifying a set of key words that have an associated measure intradocument orthogonality.

30 In an optional step, processing the assigned text may include identifying a set of synonyms for extending the search keys. Further, a subset of search keys may be selected that have a predetermined measure of correlation to the topic. The search keys may be applied to the database such as through an Internet search engine, to discover documents that are related to the search keys. Optionally, the Internet search engine may be a meta-search engine.

5           Once documents have been discovered from the database that are related to the search keys, documents may be further processed to determine their relationship to the topics associated with the search keys.

10           In the methods described herein creating an association between the at least one topic and the information retrieved from the database may include capturing a location pointer that is associated with the information retrieved from the database. Creating that association may include generating a data structure for the topic which allows storing location pointers that are associated with information retrieved from the database.

15           In another aspect, the systems and methods described herein include systems for organizing a collection of documents. Such systems may comprise an orthogonal corpus of information that is arranged according to an index of topics, a keyword generator for generating a set of keywords representative of a documents associated with a topic in the index of topics, a scoring system for processing documents within the collection of documents to associate with at least a portion of the documents a score representative of the document's association to a particular topic, and a graphical representation for depicting at least a portion of the index of topics and having respective portions of the graphical representation linked to documents associated with a respective portion of the index of topics.

20           In a further embodiment the systems of the invention may include systems for extending the content of the document. These systems can include a parser for selecting terms within the document to be extended, an orthogonal corpus of information arranged according to an index of topics, a keyword generator for generating a set of key words representative of a documents association with a topic in the index of topics, and a linking system for processing the documents within a collection of documents, to associate with at least a portion of the document, a score representative of the documents association to a particular topic and for providing the first document with links to the collection of documents for extending the content of that document.

25           Accordingly, the systems and methods described herein may leverage the electronically stored content of the World Wide Web in an intelligent and meaningful way, to provide a database of content organized under an orthogonal and hierarchical index of topics and subtopics.

5           Other objects of the invention will, in part, be obvious, and, in part, be shown from the following description of the systems and methods shown herein.

### **Brief Description of the Drawings**

10           The foregoing and other objects and advantages of the invention will be appreciated more fully from the following further description thereof, with reference to the accompanying drawing wherein;

          Figure 1 depicts a screen shot of a portion of a orthogonal corpus and a set of documents and the scores associated with those documents.

15           Figures 2A through 2B depict a dataflow diagram of one process according to the invention for processing a body of reference material for organizing a collection of documents according to a hierarchical arrangement of topics provided by the reference material;

          Figure 3 depicts one flow chart diagram of a process according to the invention;

          Figure 4 depicts one system according to the invention; and

20           Figures 5-9 depict a further practice according to the invention organizing content according to indices generated from a plurality of references.

### **Detailed Description of the Illustrated Embodiments**

25           To provide an overall understanding of the invention, certain illustrative embodiments will now be described. However, it will be understood by one of ordinary skill in the art that the systems and methods described herein can be adapted and modified to provide systems and methods suitable for other applications and that other additions and modifications can be made to the illustrated embodiments without departing from the scope hereof.

30           It is a realization of the invention that the indexing and cataloging of content on the Internet, as well as from other stores of information, may be performed by applying a process that employs an orthogonal corpus, or corpora, of information, such as an  
35           Encyclopedia. To this end, the processes described herein identify the topics discussed

5 within the corpus. The process also identifies within the corpus a set of keywords that are relevant to the topics presented in the corpus. The keywords associated with a topic may be employed to identify documents stored in another database that are related to the topic. A graphical representation of the index of topics found in the corpus may then be generated, with individual topics operating as links to these related documents. Thus, a  
10 user interested in reviewing content in the corpus related to a certain topic, may also activate a link in the graphical representation of the index to access other documents that have been identified as related to the topic of interest to the user.

Turning to Fig. 1, there is depicted a graphical user interface 10 of the type  
15 created and employed by systems according to the invention. The graphical user interface 10 represents a topic index 12, a portion of which is shown in this illustration. The topic index 12 may be a graphical representation of the table of contents of an encyclopedia, or other corpus. A user may employ the graphic interface 10 to access information that relates to the different topics listed in the index 12. Additionally, the  
20 depicted index 12 includes topics and subtopics, including subtopics of the same ancestor topic. For example, in Fig. 1, the topic Human Origins is the ancestor topic for the subtopics, The Study of Ancient Human and the Distribution of Early Hominids. A topic, or a subtopic, may be understood to include, optionally, its ancestor topics or underlying subtopics.

25 The graphical representation of the index 12 may include a hypertext link, or other linking mechanism, for each topic or subtopic in the index 12. For example, the user may activate the links, as depicted by the highlighted topic PHYSICS in Fig. 1, to retrieve a group of documents having content that is associated with the selected topic. As further depicted by Fig. 1, the system 10 may provide a display 20 such that for a  
30 selected topic or a subtopic, such as the selected topic Physics, a document 18, or a plurality of documents 18, may be presented to the user as documents associated with the topic. In the depicted embodiment, a pointer to the document, such as the title and URL 14 may be presented to the user. Additionally, an associated numerical score 16,  
35 that represents that document's association to the topic may also be presented. The

5 development of such scores 16 will be described in more detail hereinafter. Optionally,  
all the documents associated with a topic may be displayed in a window 20 of the  
system 10.

Turning now to Figures 2A and 2B, dataflow diagrams are presented that  
10 illustrate one process for creating a graphical interface, such as the interface 10 of Fig. 1.  
Specifically, Figures 2A and 2B depict a process 30 wherein a corpus, such as an  
existing published book of reference material, is processed by an orthogonal corpus  
indexing (OCI) process that extracts content signatures and topic indices from the  
corpus' content. The depicted process employs the content signatures to generate search  
15 strings for search engines to identify content associated with topics described in the  
corpus. The retrieved or discovered documents may be examined for content relevance  
and the relevant documents may be associated with topics presented in the orthogonal  
index of the corpus. Optionally, site attributes such as document type, timeliness, source  
and other such attributes may also be identified and employed to select relevant websites  
20 that may be associated with a topic in the index of the orthogonal corpus.

More specifically Fig. 2A depicts that the process 30 operates on a corpus 32 that  
may be input to the index generator 34. The index generator 34 may generate an index  
for the corpus 32 and this index may be provided to the keyword generator 48. The  
25 keyword generator 48 may produce a set of key words 52 and may be associated with  
the index 40. The process 30 continues in Figure 2B which shows the index 40 in the  
search keys 52 being applied to a search engine 54. The search engine 54 discovers  
documents from a database of content, or from a collection of databases of content 58 to  
thereby create an association between at least one of the topics of the index 40 and the  
30 information retrieved from the database 58.

The depicted corpus 32 may be any collection of information and may include,  
but is not limited to, encyclopedias, text books, dictionaries, thesauruses, atlases, maps,  
and other reference material. In one embodiment, the corpus 32 may be a published  
35 book that may be turned into or stored in an electronic format such as a conventional

5 computer datafile of text information. The corpus 32, preferably in an electronic format, may be provided to the index generator 34. The index generator 34 may process the corpus 32 to identify a hierarchical organization of a plurality of topics that appear within the corpus 32. To this end, the index generator 34 may decompose the corpus 32 to create a standard hierarchical topic orientation that is capable of assigning text content to title, headers, topics, subtopics, or any other device that may be employed for representing a section of text related to a topic, meaning, category, or some other similar abstraction.

15 US Patent 5,963,203 entitled "*Automatic index creation for a word processor*" issued to Sotomayor, Bernard, describes methods that may be employed by the index generator 34. For example, Sotomayor describes methods that enable scanning one or more documents to automatically identify key topics and phrases in a document's text, as well as methods to generate an index to those key topics. Similarly, US Patent 5,819,258 entitled "Method and apparatus for automatically generating hierarchical categories from large document collections", by Vaithyanathan, Shivakumar, Travis, Robert, and Prakash, Mayank, further describes techniques that may be employed by the index generator 34 for determining an index for a corpus. Other techniques known in the art may also be employed by the index generator 34 without departing from the invention.

25 In an alternative practice, the index generator 34 allows an operator to identify the type of corpus 32 being input into the index generator 34. For example, the index generator 34 may present an interface to the operator that allows the operator to identify whether the corpus being presented comprises an encyclopedia, a dictionary, a textbook, or another known type of reference document. Additionally, the index generator 34 may allow the operator to identify whether the corpus 32 includes a table of contents, an index, chapter heading, or any other representation of the different topics contained within the corpus. In this embodiment, the user may identify, for example, that the corpus 32 comprises an encyclopedia and that the encyclopedia includes a table of contents that is representative of the index of orthogonal topics maintained within the

5 encyclopedia. In this embodiment, the index generator 34 may process the presented corpus 32 to identify the table of contents for the encyclopedia. This table of contents, in one embodiment, may be formatted into an HTML document that presents the table of contents in an organized format that emphasizes the topics, subtopics and other hierarchical structure of the table of contents. In one process the index generator 34  
10 processes the notation for the table of contents, such as the topic numbering employed, to identify which topics are understood as parent topics and which are understood as main topics and which are understood as subtopics. In a further optional embodiment, the index generator 34 may present the generated index with the orthogonal corpus 40 to the operator to allow the operator to edit or amend the generated index for the  
15 orthogonal corpus 40.

As shown in Figure 3B once the index generator 34 has processed the corpus 32 the index generator 34 may present the index 40 for the corpus 32 to the keyword generator 34. As depicted in Figure 2A, the index 40 may comprise a hierarchical  
20 representation of the orthogonal topics maintained within the corpus 32. This hierarchical representation may include primary topics, such as the depicted topic 38 and a plurality of subtopics 42 that are associated with the primary topic 38.

The keyword generator 48 in one embodiment operates to identify sections of text  
25 of the corpus 32 to be associated with the different topics and subtopics of the index 40. Continuing with the above example, in those practices where the index 40 is generated from the table of contents for the corpus 32, the keyword generator 48 may identify those pages that contain information associated with a topic presented within the index 40. For example, the keyword generator 48 may process the table of content for the  
30 corpus 32 to identify a page number associated with a topic, such as the topic 40 and may analyze the page associated with topic 40 to identify that portion of the page that may be associated with the topic 40. In one embodiment, where headings are presented within the corpus 32, the keyword generator 48 may analyze the page associated with the topic 40 to identify a heading that is representative of the beginning of the  
35 presentation within the corpus 32 of information that is associated with topic 40. For

5 example, the keyword generator 48 may identify a section of text within the associated  
page that contains the information associated with topic 40 and that is presented in a  
type font and font size that is representative of a heading. In a subsequent step the  
keyword generator 48 may identify the location of the heading for the subsequent topic  
10 44 that indicates the beginning of content related to the new topic. The keyword  
generator 48 may identify the content that is delimited by the heading 40 and 44 and  
associate that content as content related to the topic 40.

15 Once the portion of the corpus 32 that is to be associated with the topic 40 is  
identified, the keyword generator 48 may process this assigned portion of text to  
generate a plurality of search keys, each of which may be representative of a search  
string for selecting information from a database.

20 In one embodiment, the system 10 employs the orthogonal construction of the  
corpus for algorithmic identification of keywords in each topic that distinguish that topic  
from its sibling, cousin, ancestor, or descendent topics. Accordingly, the systems  
described herein may create a set of keywords for a topic that identifies a document  
associated with a topic and that are keywords which may act to distinguish documents  
associated with one topic, from documents associated with another topic. For example,  
the system 10 may employ processes that identify keywords that are associated strongly  
25 with a particular topic. Techniques for creating keywords will be understood from  
Deerwester, S., Dumais, S.T., Landauer, T.K. Furnas, G.W. and Harshman, R.A. (1990).  
"Indexing by latent semantic analysis." *Journal of the Society for Information Science*,  
41(6), 391-407. Additionally, the system 10 may identify other keywords that act to  
disassociate a document from one or more other topics. These keywords may be  
30 employed by the system 10 to numerically score over an underlying pool of documents.

35 The system 10 may employ scoring methods that may utilize traditional  
information retrieval techniques including the use of synonyms, stemming, frequency,  
proximity, stop words, hyponyms, and synonyms. If, as in most large document  
collections, it is not practical for all documents to be individually scored against the



5 keywords, then a subset of search words is selected to identify candidate documents for  
scoring. Keyword and search terms are identified based on a numerical method that  
apportions words among topics. The goal is for the keywords and search terms to  
10 identify individual blocks of text as found at the nodes of the orthogonal corpus topic  
hierarchy. In an ideal sense, the keywords would be partitioned across the hierarchical  
tree nodes, with each word occurring in only one corpus topic. In addition to word rarity  
among corpus topics, rarity in the underlying document pool may contribute to a word  
being identified as a keyword or search term for a given topic. For example, a keyword  
15 occurring in only one node and only once on the Web, would be a top candidate as a  
keyword and search term.

20 The keyword generator 48 may present as an output, a set of keywords 52 each  
of which may be associated with a topic or subtopic in the index 40. As described  
above, these keywords may be employed to act to distinguish documents associated with  
one topic from documents associated with another topic. Accordingly, as depicted in  
Figure 2B the search keys 52 and associated topics on the index 40 may be presented to  
the search engine 54 for retrieving information from a database or databases of content  
58. To this end, the process 30 applies the search keys to the database 58 to retrieve  
information from the database 58. In one practice as will be described in more detail  
hereinafter, an optional step in process 30 is performed wherein the search keys 52 are  
25 processed to identify a subset of search keys that may be employed for generating search  
queries to one or more search engines, such as Internet search engines, to discover a set  
of documents which are relevant to the topic of interest. Each of the resulting  
documents may be examined in a subsequent step to determine the relevance of the  
content contained within the index. The relevance may be scored, as further described  
30 below, for identifying the relevance of that document, and the score may be employed  
for ordering the sequence in which content is listed as being relevant to a particular  
document.

35 Once the discovered documents are scored for relevance, the process 30 may  
associate portions of the discovered documents to associated topics within the index 38.

5 In a practice wherein the database 58 includes links to URL's for websites, the process  
30 may create a web database that contains website information such as URL's, types,  
dates, topics, contents, size and editor notes that are inserted or updated in the database  
from time to time. Information about the corpus 32 that has been processed, such as the  
10 publisher, the ISBN, and other types of information needed to purchase the book through  
an online transaction may also be stored. The search engine may then provide a  
navigation tool that comprises the HTML representation of the index 38 wherein topics  
and subtopics within the index 38 link to URL's of web content identified as being  
related to the topic or subtopic selected by the user. Optionally, in certain embodiments,  
15 the topics and subtopics may also include links to portions of the corpus 32 that are  
related to the topic selected by the user. In this way, a user may select a topic presented  
by the corpus 32 in view of the information presented by the corpus 32 and related  
information stored on the World Wide Web. In other embodiments, other techniques are  
employed for semantic processing and for determining a topic that may be associated  
20 with a portion of text within the corpus.

20 The data flow diagram depicted in Figs. 2A and 2B may be implemented in a  
data processing process wherein a data processing program processes the corpus and  
generates an index that links topics in the corpus to information from a data sources,  
such as the Internet. Turning now to Fig. 3, a flow chart illustration of one such process  
25 is depicted. Specifically, Fig. 3 depicts a process 70 for extending a corpus by  
identifying topics covered by that corpus and employing information stored in the  
corpus and related to the topics to identify information in a database that is also related  
to the topic. The process 70 also generates an optional graphical user interface, such as  
the interface depicted in Fig. 1, that includes links for topics listed in the index, and that  
30 may be employed by a user to access the information associated with the listed topics.

35 The process 70 begins with the act 72 of identifying a corpus that is to be  
extended, such as by selecting a publication that contains reference material. In step 72,  
the process 70 transforms, or casts, the corpus into a normal form for processing. In one  
practice, this involves decomposing the document format of the corpus into a standard

5 hierarchical topic orientation with a mechanism for assigning text content to title, headers, topics, and sub-topics. Optionally, stop words, such as the common words "and", "them", and "within", are identifies and removed during normalization.

10 After normalization, the process 70 proceeds to step 78 wherein the corpus is processed to identify which portions of the corpus relate to which topic. In one practice wherein the corpus includes a table of contents, the process 70 analyzes the document format of the corpus to locate within the text headings associated with the different topics. For example, as described in the above cited publication US Patent 5,963,203 entitled "*Automatic index creation for a word processor*", header information set off by  
15 HTML tags may be identified to find indicia of topic entries in the document being processed. However, any technique for processing a document to identify the sections of text related to a topic may be applied, including other techniques for analyzing the mark up form language of the document.

20 Proceeding to step 80, the process 70 analyzes the topics to identify a signature that may be understood as representative of the semantics of the topic. In one practice, the process 70 creates a word map per topic and subtopic. To this end the process 70 in step 80 may create a summary representation of the words in the text based on the number of, location of, and proximity of words within each topic and sub-topic. Other  
25 factors may be employed, or substituted for these. Statistics are maintained on different parts of the document structure such as titles, headings, paragraphs, sentences, and image.

30 Table I depicts that several topics may be identified within the corpus. For example, Table I depicts that the processed corpus includes the topics Archaeology, Argentina, Arithmetic, Art, and Astronomy.

<u>Topic</u>
Archaeology

**Argentina**

**Arithmetic**

**Art**

**Astronomy**

**TABLE I**

The process 70 in one practice may then determine for a given topic, the word count for the words that appear within the portion of text, or other content, associated with the respective topic. This is depicted in Table II, that shows an example of the word count, with stop words removed, for words that appear in the portion of the corpus related to the topic "Astronomy."

**Word counts in Topic Astronomy**

<b>Word</b>	<b>Count</b>
<b>actual</b>	<b>1</b>
<b>ad</b>	<b>1</b>
<b>adopted</b>	<b>1</b>
<b>advances</b>	<b>2</b>
<b>ancient</b>	<b>3</b>
<b>application</b>	<b>2</b>
<b>assigning</b>	<b>1</b>
<b>astronomer</b>	<b>2</b>
<b>astronomers</b>	<b>2</b>
<b>astronomical</b>	<b>5</b>
<b>Astronomy</b>	<b>12</b>
<b>astronomy</b>	<b>11</b>
<b>Astrophysicists</b>	<b>1</b>
<b>astrophysics</b>	<b>1</b>

**TABLE II**

In process 70, after the word count, and other statistics are determined, signatures are generated using orthogonalization. For example, in one practice, given the word counts or word maps for all or a selected subset of topics simultaneously, the process 70

5 assigns a weight based on word count to each word within each topic or subtopic.  
 Where using word counts the weight may be defined as the count. When using the word  
 map, the weight of a word in a topic or subtopic may be assigned by an intra-document  
 scoring function. Any suitable technique may be employed for performing intra-  
 document scoring. These signatures may be edited or cleaned manually to enhance the  
 10 topical relevance and precision of the subsequent search and scoring process. Table III  
 depicts an example signature for the topic "Astronomy."

Signature for topic Astronomy	
Word	Count
Astronomy	23
earth	9
bodies	5
Astronomical	5
universe	4
celestial	3
circle	3
Observational	3
sky	3
Stars	3
Ancient	3
Daily	3
Heavenly	2
Astronomers	2
planet	2
relative	2
moved	2
heavens	2

TABLE III

After determining a signature, the process 70 may perform the optional step, step  
 82, of applying synonym Groups. In this optional step, the process 70 extends the

5 signatures with synonym groups. To this end, words are replaced by groups of word substitutes having similar or identical meaning. Table IV depicts such an extension.

Astronomy Signature Post Synonym Reduction	
Word	Count
Astronom	30
earth	9
bodies	5
universe	4
celestial	3
circle	3
Observational	3
sky	3
stars	3
ancient	3
daily	3
Heavenly	2
planet	2
relative	2
moved	2
heavens	2

TABLE IV

30 After step 82, the process 70 may proceed to step 84, wherein the process reduces the signature to Keyword sets, optionally tailored for the search. The set of documents to be scored against a topic is preferably identified and manageable in size. The web for example is a large a document set to collect up and score against all web documents. Accordingly, in one practice traditional large scale search engines, such a

35 Lycos and Alta Vista, may be used to identify a set of candidate relevant documents using a keyword set for search. Which subset of the Signature and synonym groups is included in the Keyword set may be determined based on a variety of measures

including corpus document word count of the word and general frequency of the word. An example is presented in TABLE V.

Word	Count
Astronomy	30
Earth	9
Bodies	5
Universe	4
Celestial	3
Circle	3
Observational	3
Sky	3
Stars	3
Ancient	3
Daily	3

TABLE V

The keyword set may be applied to a search mechanism to pull in multiple discovered documents based on the keyword set. This may occur in step 88. For example, the query Find: Astronomy or Astronomical or Astronomers or earth or bodies or universe or celestial or circle or Observational or sky or stars or ancient or daily; may be generated from the keyword set and applied to the search mechanism to discover documents related to the selected topic.

After the step 88, the process 70 may proceed to step 90 for scoring of the discovered documents. The many discovered documents returned from a search function may be assigned individual scores against the corresponding corpus topics and subtopics. Scoring may be based on multiple tunable metrics and rules including functions over the word count or word\_map data structures. The score of topical overlap between two documents as a baseline is measured as a dot product of word counts or word frequencies within those documents).

## Scoring example

Word	Count in Astronomy	Count in discovered document	Score Contribution
Astronom	30	2	0.555556
earth	9	1	0.083333
bodies	5		
universe	4	5	0.185185
celestial	3		
circle	3	2	0.055556
Observational	3		
sky	3		
stars	3		
ancient	3	1	0.027778
daily	3	2	0.055556
Heavenly	2		
planet	2		
relative	2		
moved	2	1	0.018519
heavens	2	1	0.018519

After step 90, the process 70 proceeds to optional step 92, wherein the topic hierarchy and set of associated documents may be presented directly through an HTML or graphical user interface, such as the interface depicted in Fig. 1. Alternatively, content may be delivered through software API's (application program interfaces) to allow integration of output content with other content. Content may be navigated by walking the directory tree structure, or by keyword searching over the directory structure trees, corpus content, or discovered document content. Search results may point to topic paths or discovered documents,

Fig. 4 depicts one embodiment of the system 100 according to the invention. Specifically Figure 4 depicts a functional block diagram that shows a system 100 that allows a surfer 102 to access a user interface 104 that couples to a database system 108. The database system 108 further couples to an OCI processor 112 that accesses a database of corpora 114 and a plurality of search engines 118. The database system 108 further couples to an application programming interface access layer 120 and through the API 120 can access a portal/search client 122. Additionally, the API 120 may also couple to a scoring mechanism 124.



5           More particularly, Figure 4 depicts that a user 102 such as an Internet user may access a user interface 104, that may be similar to the user interface depicted in Figure 1. As shown in Figure 1 the user interface 10 may present to the user 102 a list of topics 112. The user 102 may select a topic from the index 112. As described with reference to Figure 1, the selection of a link directs the user interface 104 to retrieve information from the database system 108. The database system 108 processes the users request  
10           from user 102 for information related to the selected topic.

          The database system 108 may be any suitable database system, including the commercially available Microsoft Access database, and can be a local or distributed  
15           database system. The design and development of suitable database systems are described in McGovern et al., *A Guide To Sybase and SQL Server*, Addison-Wesley (1993). The database 108 can be supported by any suitable persistent data memory, such as a hard disk drive, RAID system, tape drive system, floppy diskette, or any other suitable system.

20           As further depicted by Figure 4 that database system 108 may communicate with the OCI mechanism 112. The OCI mechanism 112 may be, in one embodiment, a computer process capable of implementing a process such as process 70 depicted in Figure 3. The OCI mechanism can be realized as a software component operating on a  
25           conventional data processing system such as a Unix workstation. In that embodiment, the OCI mechanism can be implemented as a C language computer program, or a computer program written in any high level language including C++, Fortran, Java or basic. Techniques for high level programming are known, and set forth in, for example, Stephen G. Kochan, *Programming in C*, Hayden Publishing (1983). Accordingly, the  
30           OCI mechanism 112 may be employed by a system administrator to process corpora stored within the database 114. As discussed with reference to Figure 3, the processed corpora results in a graphical user interface that may be stored within the database mechanism 108 and accessed by the user 102 through the topic navigator 104. Additionally, the OCI mechanism 112 may generate for the processed corpora of  
35           database 114 a set of links or pointers to content that corresponds with different listed

5 topics within the index of the processed corpora. The OCI mechanism 112 may also store these associated links within the database system 108.

10 To this end, the OCI mechanism 112 may couple to one or more search engines 118 that allow the OCI mechanism 112 to retrieve content from a database source. In the depicted embodiment of Figure 4, the database source that search engines 118 access is the World Wide Web 106. In this embodiment, the user interface 104 also couples to the World Wide Web 106 so that links activated by the user that relate to URL's of content stored on the World Wide Web 106 may be directly accessed by the user 102 through the user interface 104 through the connection between the user interface 104 and the World Wide Web 106. Figure 4 further depicts that the database 108 communicated with an API layer 120. As shown in Figure 4 the API layer sits between the portal search client 122 and the database system 108 and also sits between the scoring mechanism 124 and the database system 108. Accordingly, a portal search client such as the Yahoo site may access the database system 108 through the API layer to provide users with access to an index linked to content on the World Wide Web.

25 Similarly, Figure 4 depicts the scoring mechanism 124. The scoring mechanism 124 may be a computer process that accesses the database system 108 through the API 120. The scoring mechanism may perform data mining for identifying topics that are to be associated with different websites. In this way, the database system 108 may be employed for categorizing web sites according to their content. Thus, the system 100 depicted in Figure 4 provides a system for categorizing information stored on the World Wide Web. The system described in Fig. 4 may operate on any suitable computer hardware, such as PC compatible computer systems, Sun workstations, or any other suitable hardware. The list of topics and the associated documents, or links to documents may then be stored in a relational database, or any suitable database with proper indexing for allowing rapid accessing of the data stored therein.

5           Once the system is operating, the system may be employed to provide a set of tools, such as that may operate as stand alone applications for single users, or that may be tools provided as client/server programs over a network. The tools may be provided as a collection of functions incorporated into an integrated research tool, or may co-exist as individual functions in a separate application.

10           Further embodiment, the systems and methods described herein may be employed for organizing a plurality of corpora into an indexed format that may be presented as a graphical user interface for a user to allow a user to access information related to the contents of a plurality of corpora. For example, figures 5 through 9 depict  
15           the operation of a system that processes a plurality of text, such as reference texts. Accordingly, the system may be employed for the automatic creation of a topically organized book catalog, such as a catalog of reference books, with navigation, search, click through to external documents such as web documents, with information  
20           purchasing interfaces also. For example, figure 5 depicts a graphical user interface that presents to a user a plurality of topics each having a set of books within the topic. For example, the figure 5 depicts a topic reference that includes a set of encyclopedias and dictionaries within that reference. By activating the reference link, the user may be  
25           presented with the user interface shown in figure 6. Figure 6, the individual references presented under the reference topic of figure 5 are outlined for the user allowing the user to select what type of reference the user would like to view.

          For example, the user may select from encyclopedias, dictionaries, academic and learned society publications and other such publications. After making a selection  
30           figure 6, the user may be presented with the different books under each category. The example presented in figures 5 through 9 shows that upon activating the link for encyclopedias, the user is presented with the different encyclopedias that have been processed by the system according to the invention. Upon selecting a link, such as the link for the Encyclopedia Britannica, the user may be presented with the interface shown in figure 8 that lists the different topics covered by the Encyclopedia Britannica. At this  
35           level, the process now proceeds as described above, with reference to figures 1 through 4

5 wherein the individual topics maintained within the Encyclopedia Britannica may be employed for accessing contact, such as web contact particularly associated with the individual topics.

10 Those skilled in the art will know or be able to ascertain using no more than routine experimentation, many equivalents to the embodiments and practices described herein. For example, the systems and methods described herein may be employed for providing encyclopedia (i.e., corpus) extender. An encyclopedia (as an archetype example of an orthogonal corpus) may be automatically extended by application of the systems and methods described above, to include links into the World Wide Web, or  
15 other database, via searching or meta-searching over the Web. The breadth and depth of the corpus enables a high quality, high coverage database of web links, with the web links organized according to the location in the topic hierarchy whose text was used to generate them. Such links may provide geographical maps, histories of topics of interest, access to theses and other types of information. Other applications include web book companions wherein the system processes a book, including a fictional work,  
20 a non-fictional work, or a reference book, through this system will allow automated construction of topical web sites as Web Companions to individual books. For example, a book such as the Hunt for Red October may be processed by the systems described herein to create links into the Web to documents associated with concepts from the book, such as links to the Navy Submarine division, links to topographic maps of the  
25 ocean floor, links to Russian Naval History, and other similar links.

A search engine extension may be provided by accessing the database 108 through the API. Thus a user may do a search on a web search engine, they may want to refine their search or get a second search opinion. Given a broad topic database such as  
30 that created in the Encyclopedia Extender application described above, refinement of a user's intended topic is enabled – through keyword-based narrowing, web link browsing, and display of proximal or correlated topics in the corpus topic hierarchy. For e-commerce, the systems described herein book/article browser/seller. Browsing over the  
35 topic hierarchy may provide indexes into books or articles for sale.

5 Additional applications can include a user Interface. The user interface allows users to view Web links through the topic hierarchies defined by the corpus. The topic hierarchy on the left lists the topics as per the corpus. The user may select keywords from the corpus outline, or from provided sample text inside the corpus documents, to better focus and score the topic. Users may augment the search terms or keywords with their own keywords or selected synonyms to more specifically tailor a concept to a need. Searching across the corpus or across the referenced links may include synonyms, stemming, frequency, proximity, stop words, hyponyms, and synonyms.

10 Additionally, authoring toolkits may be provided that allow publishers, editors, and authors to create corpus extensions and associated applications. For example, the systems and methods described herein may be employed to create development kits that publishers may use to index a book and create a web site that acts as the book companion described above.

15 It may be noted that human oversight or auditing of the document scoring and database may be done in order to augment the purely automated document selection. This may be done on a sampling basis for quality control. Different levels of sensitivity to content or product price points may be implied by different levels of human quality control. Moreover, it will be noted that the system described above has been done so with reference to documents stored on the Web. However, it will be understood by those of ordinary skill in the art, that the Web is being used here as a metaphor for any electronic document archive, and the systems and methods described herein are not limited to the Web.

20 Accordingly, it will be understood that the invention is not to be limited to the embodiments disclosed herein, that other applications, such as information mining may be practiced with the systems and methods described herein, and that the invention is to be understood by the following claims which are to be interpreted as broadly as allowed under the law.

5

We claim:

1. A method for processing a body of reference material to generate a directory for accessing information from a database, comprising

10

processing the body of reference material to identify a hierarchical organization of a plurality of topics,

associating with at least one of said topics a portion of the reference material,

processing said assigned portion of the reference material to generate a plurality of search keys representative of search strings for selecting information from the database,

15

applying the search keys to the database to retrieve information from the database, and

creating an association between the at least one topic and the information retrieved from the database.

20

2. A method according to claim 1, further comprising

creating a graphical interface representative of the identified hierarchical organization of a plurality of topics for allowing a user to access information retrieved from the database and having an association with a topic.

25

3. A method according to claim 1, wherein processing the body of reference material includes processing a body of reference material selected from the group consisting of an encyclopedia, a dictionary, a text book, a novel, a newspaper, a web site, and [ Henry what else are we interested in processing?].

30

4. A method according to claim 1, wherein processing the body of reference material includes processing the body of reference material to identify a hierarchical organization of a plurality of substantially orthogonal topics.

35

5. A method according to claim 4, wherein processing the body of reference

5 material includes identifying a table of contents for the body of reference material.

6. A method according to claim 4, wherein processing the body of reference material includes identifying chapter headings.

10 7. A method according to claim 4, wherein processing the body of reference material includes identifying definition entries in a dictionary.

15 8. A method according to claim 4, wherein processing the body of reference material includes normalizing the identified hierarchical organization of a plurality of topics. [ Henry - here we are interested in any things we do to process the hierarchical organization to make the topics more orthogonal].

20 9. A method according to claim 1, wherein processing the assigned text includes generating a word map representative of a statistical analysis of words contained in the assigned text.

25 10. A method according to claim 9, wherein generating the word map includes performing a word count process for determining word frequency of a word within the assigned text, and for employing the word frequency for determining relevance of a word to the associated topic.

30 11. A method according to claim 1, wherein processing the assigned text includes processing the assigned text to identify a set of keywords having an associated measure of intra document orthogonality.

12. A method according to claim 1, wherein processing the assigned text includes identifying a set of synonyms for extending the search keys.

35 13. A method according to claim 1, wherein processing the assigned text includes

5                    selecting a subset of the search keys having a predetermined measure of correlation to the topic.

10           14.    A method according to claim 1, wherein applying the search keys to the database includes applying the search keys to an Internet search engine to discover documents related to the search keys.

15           15.    A method according to claim 1, wherein applying the search keys to the database includes applying the search keys to a meta-search engine to discover documents related to the search keys.

                 16.    A method according to claims 14 or 15, wherein applying the search keys to the database includes processing text of the discovered documents to identify documents related to the topics associated with the search keys.

20           17.    A method according to claim 1, wherein creating an association between the at least one topic and the information retrieved from the database includes capturing a location pointer associated with the information retrieved from the database.

25           18.    A method according to claim 1, wherein creating an association between the at least one topic and the information retrieved from the database includes creating a data structure for the topic for storing location pointers associated with information retrieved from the database.

30           19. A system for organizing a collection of documents, comprising  
                 an orthogonal corpus of information arranged according to an index of topics,  
                 a keyword generator for generating a set of keywords representative of a  
document's association with a topic in the index of topics,  
                 a scoring system for processing documents within the collection of documents, to  
associate with at least a portion of the documents, a score representative of the  
35           document's association to a particular topic, and



- 5                   a graphical representation for depicting at least a portion of the index of topic and  
having respective portions of the graphical representation linked to documents  
associated with a respective portion of the index of topics.
20. A system for extending the content of a first document, comprising
- 10                   a parser for selecting terms with said first document to be extended,  
                  an orthogonal corpus of information arranged according to the index of topics,  
                  a keyword generator for generating a set of keywords representative of a  
document's association with a topic in the index of topics, and
- a linking system for processing documents within a collection of documents, to
- 15                   associate with at least a portion of the documents, a score representative of the  
document's association to a particular topic, and for providing said first document with  
links into said collection of documents for extending the content of said document.

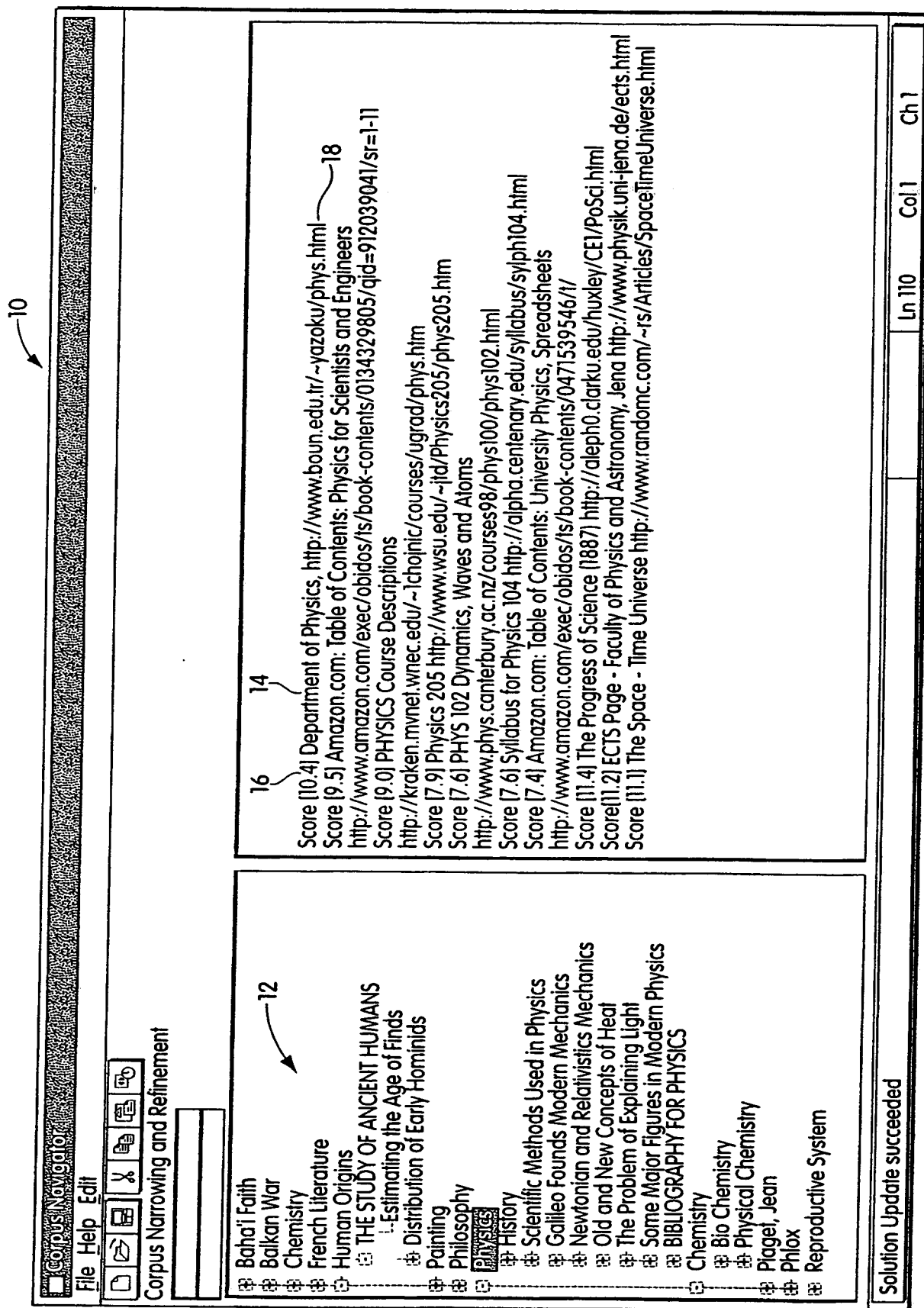


Fig. 1

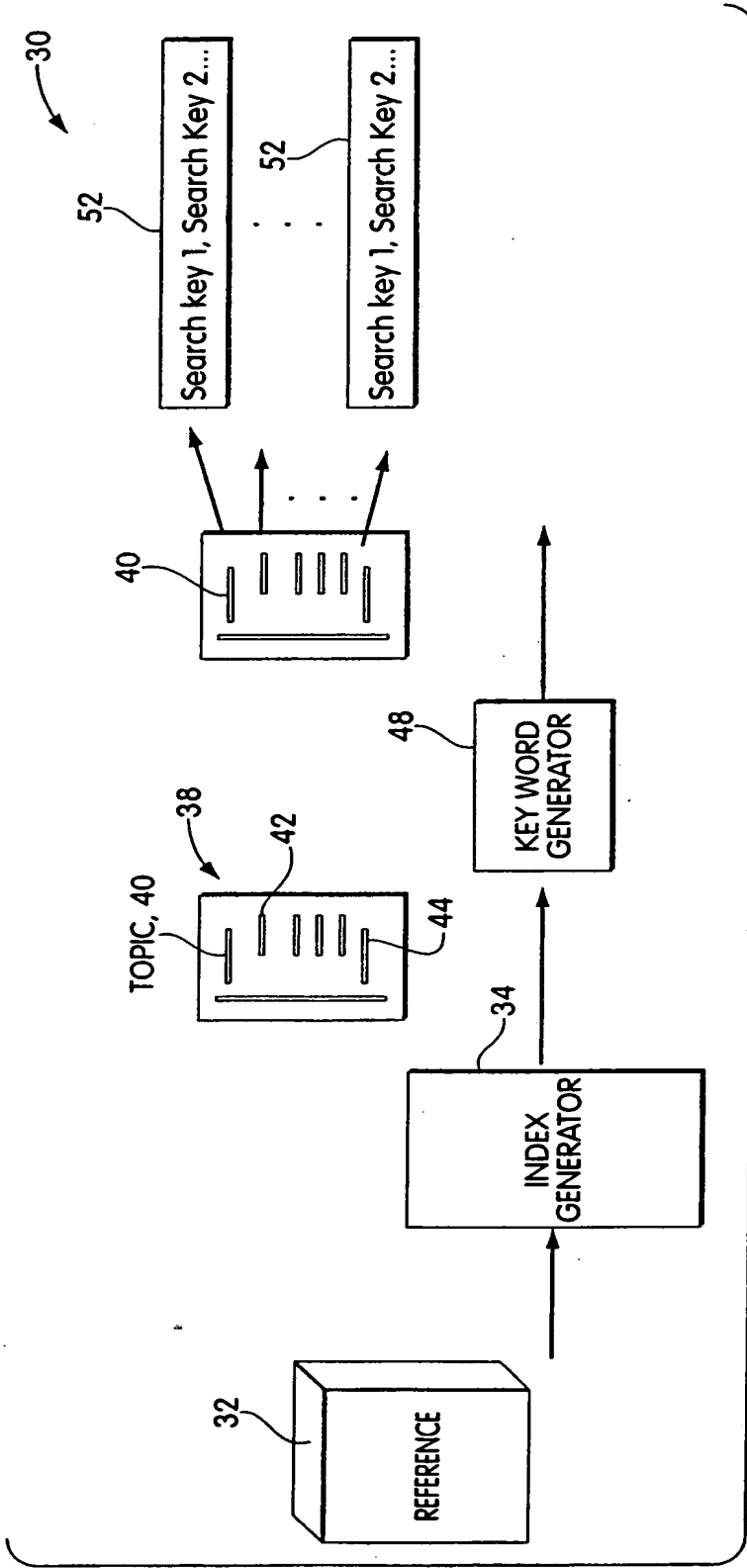


Fig. 2A

3/10

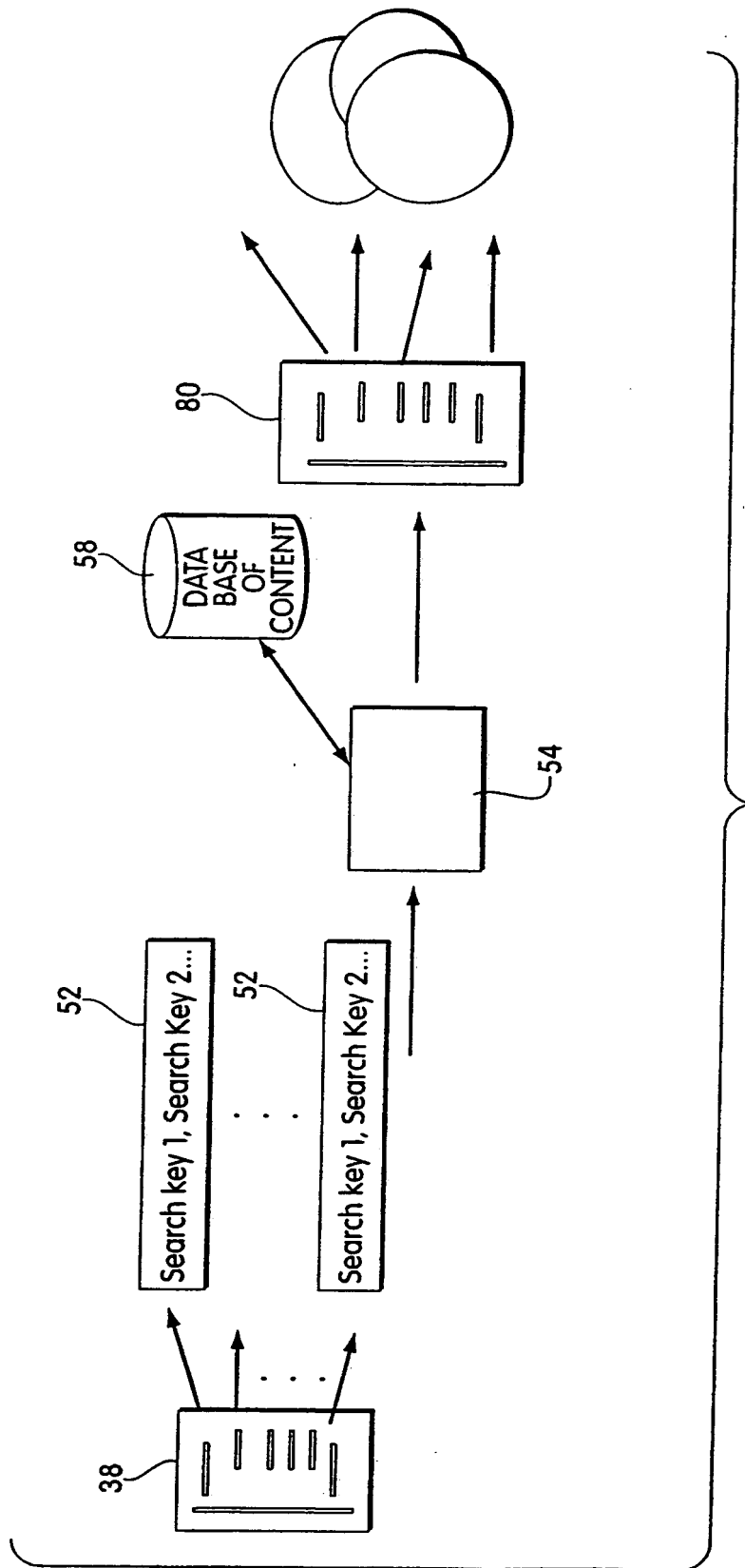


Fig. 2B

4/10

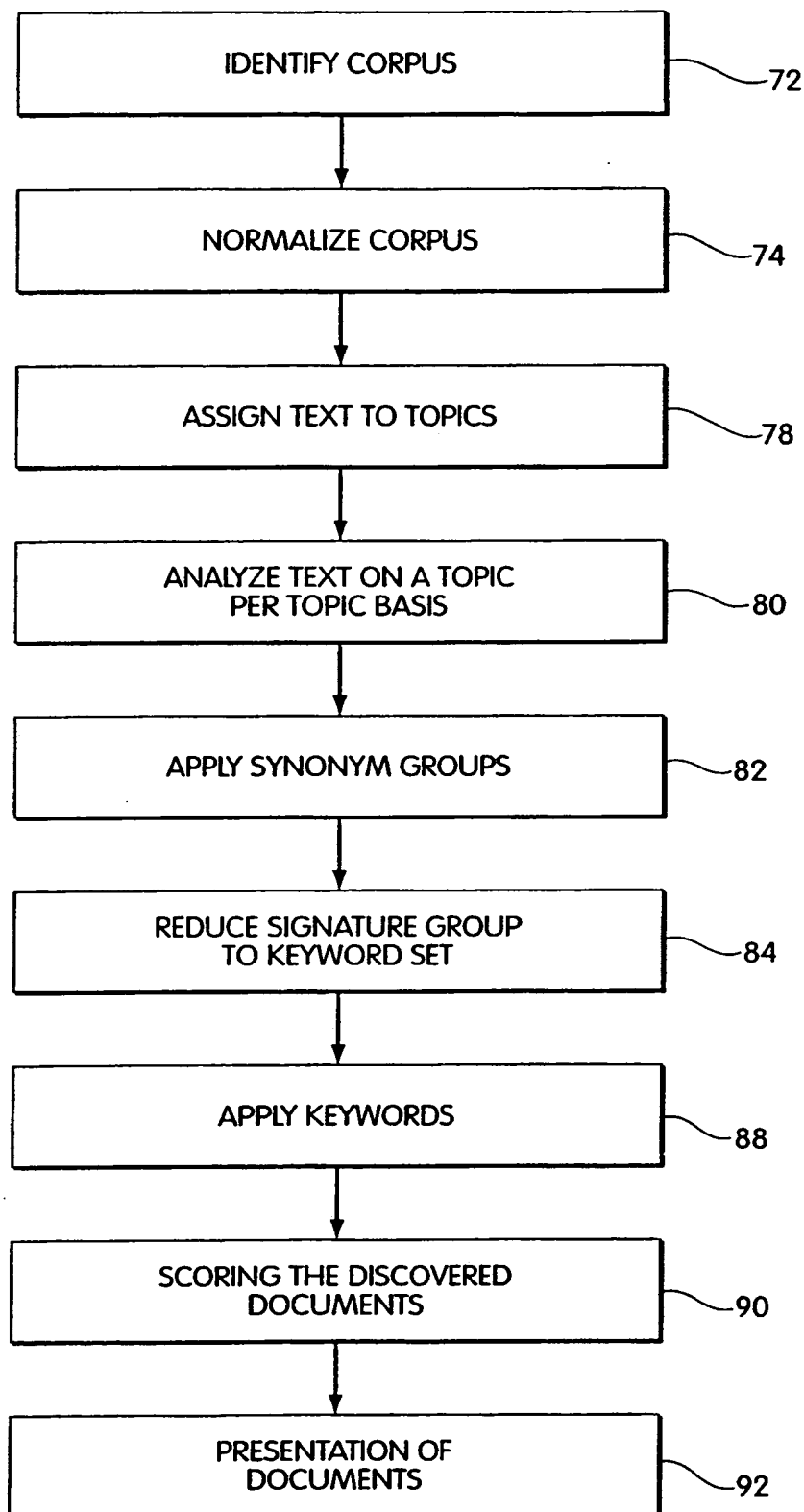


Fig. 3

5/10

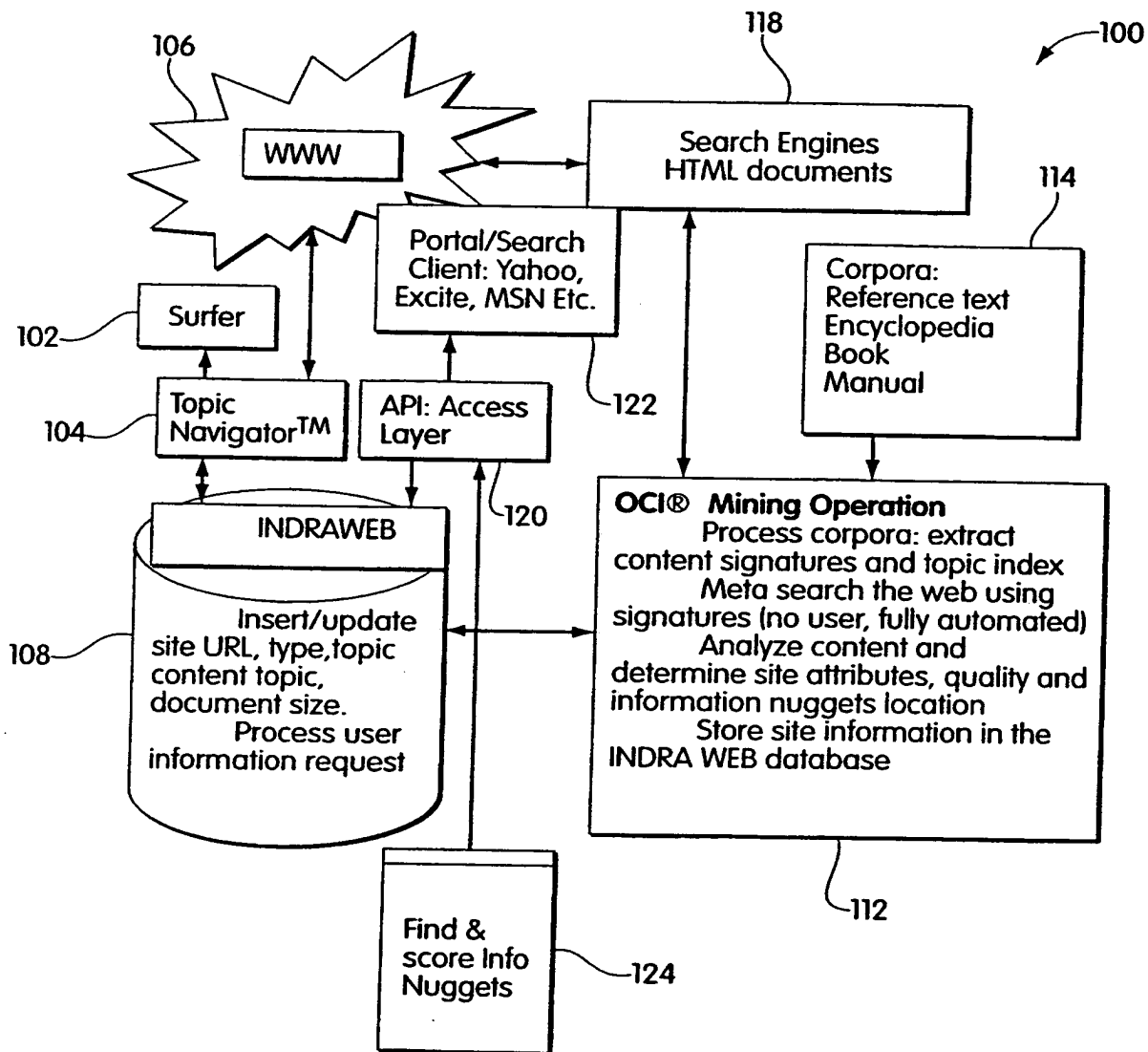


Fig. 4

6/10

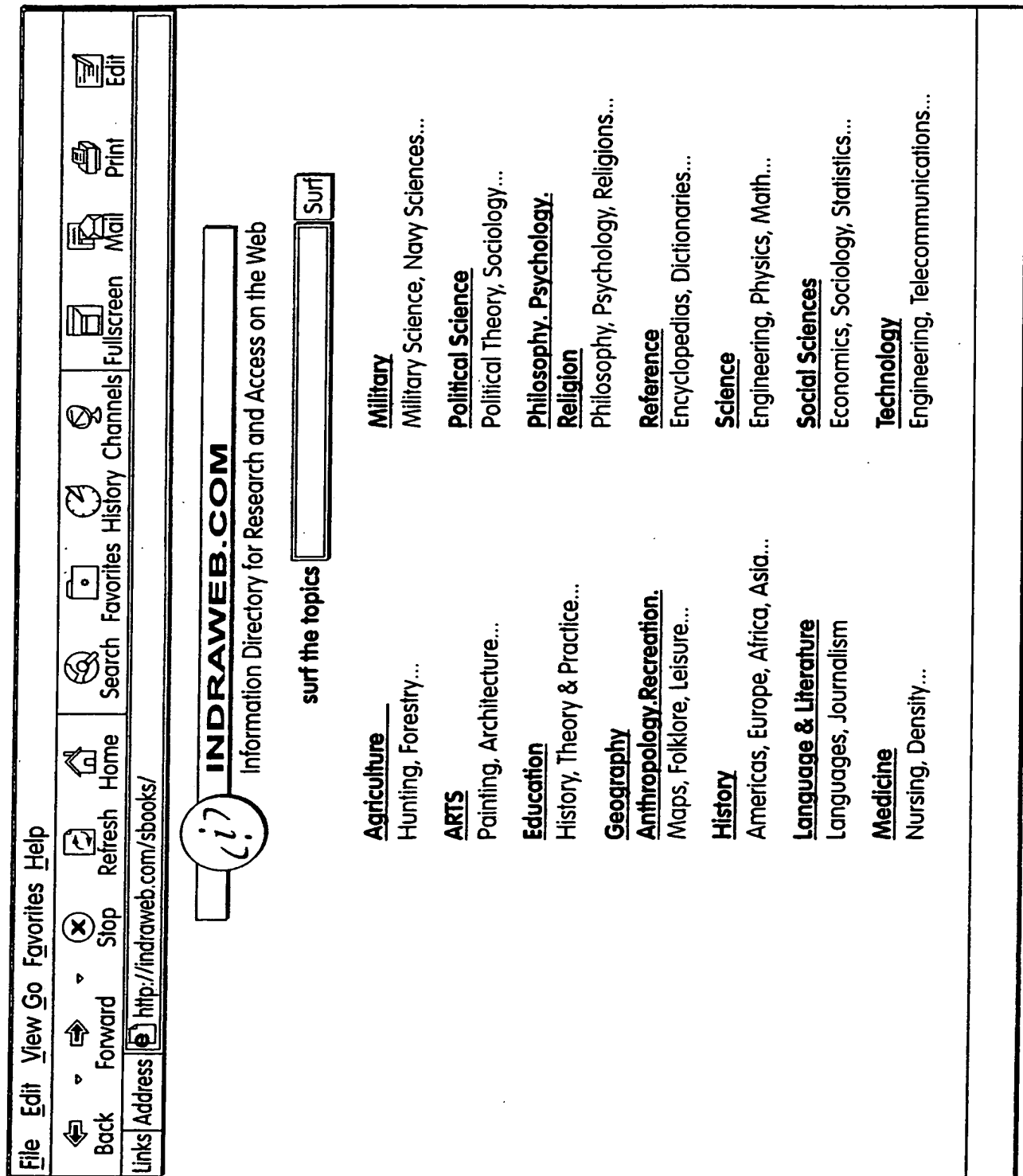


Fig. 5

7/10

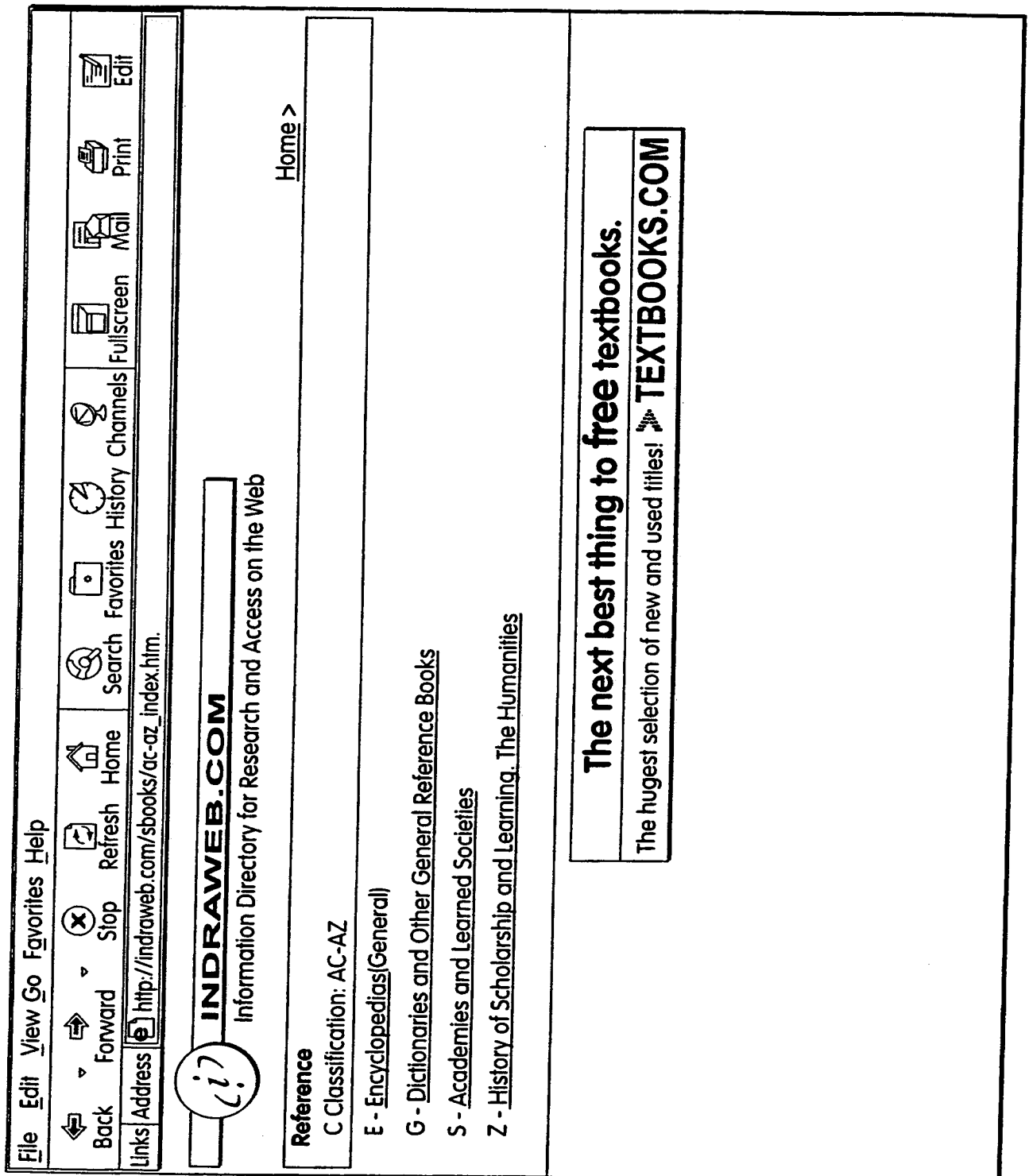


Fig. 6



8/10

File Edit View Go Favorites Help				
Back	Forward	Stop	Refresh	Home
Search	Favorites	History	Channels	Fullscreen
Links	Address	http://indraweb.com/books/ae.htm		
Home > Reference >				
Encyclopedias (General)				
Omne bonum : a fourteenth-century encyclopedia of universal knowledge : British Library MSS Royal 6 E VI-6 E VII		1996	AE2.L463 S26	H. Miller Publishers
Academic American encyclopedia.		1998	AE5 .A23	Grolier Inc.
Encyclopedia Britannica		1997	AE5 .B	Routledge
The Cambridge paperback encyclopedia		1995	AE5 .C36	Cambridge University Press
Cambridge paperback encyclopedia		1993	AE5 .C36	Cambridge University Press
Collier's encyclopedia : with bibliography and index		1997	AE5 .C683	Collier's
The Encyclopedia Americana.		1999	AE5 .E333	Grolier Incorporated
Encyclopedia of knowledge		1991	AE5 .E478	Grolier
Family encyclopedia.		1997	AE5 .F36	Oxford University Press
Funk & Wagnalls new encyclopedia		1996	AE5 .F83	Funk & Wagnalls
The Macmillan compact encyclopedia		1991	AE5 .M24	Macmillan
Merit students encyclopedia		1993	AE5 .M38	Macmillan Educational Co.
New standard encyclopedia.		1998	AE5 .N64	Ferguson Pub. Co.
Oxford illustrated encyclopedia		1993	AE5 .O94	Oxford University Pre
Webster's family encyclopedia.		1992	AE5 .W35	Merriam-Webster, Inc
The World Book encyclopedia.		1999	AE5 .W55	World Book
The World Book encyclopedia of people and places.		1998	AE5 .W563	World Book

Fig. 7

9/10

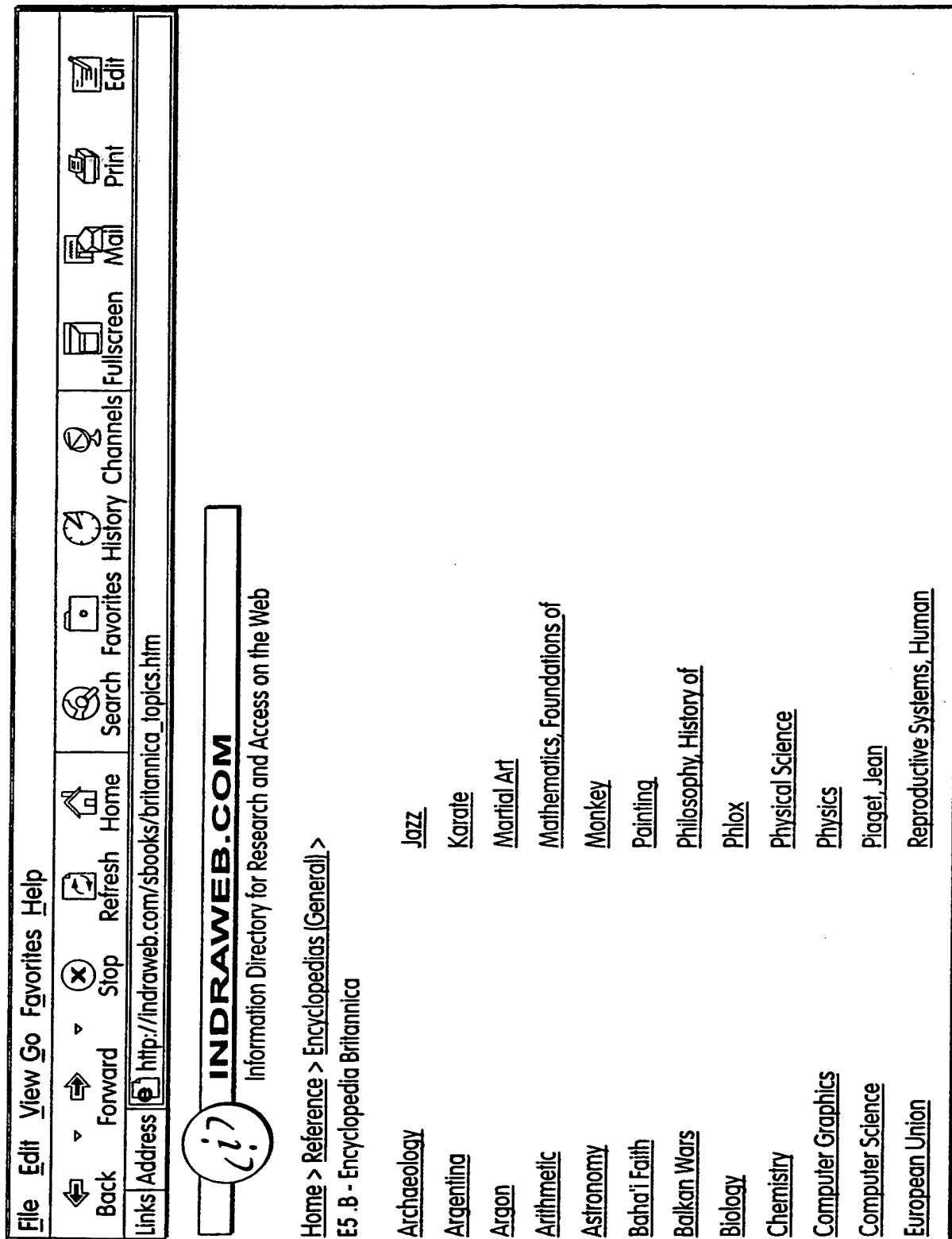


Fig. 8

10/10

File Edit View Go Favorites Help

Back Forward Stop Refresh Home Search Favorites History Channels Fullscreen Mail Print Edit

Address <http://indraweb.com/sbooks/astromy.htm>

Home > Reference > Encyclopedias (General) > Encyclopedia Britannica > **E6: Britannica Britannica Astronomy**

**Search:** astronomy earth bodies astronomical universe celestial circle observational sky stars ancient daily heavenly astronomers planet relative moved heavens ray ptolemaic copernicus revolution recognized electronic advances position rad (Any word)

**Found:** 277 document(s) on The Web Sort: S

**Imagine the Universal Dictionary**  
Please allow the whole page to load before you start searching for an entry. Otherwise, errors will occur. [A B C D E F G H I J K L M N O P Q R S T U V W X Y Z]

Found by: Infoseek, WebCrawler  
<http://imagine.gsfc.nasa.gov/docs/dictionary.html>

**APOD Glossary**  
Astronomy Picture of the Day Glossary Cepheid: A pulsating variable star. This type of star undergoes a rhythmic pulsation as indicated by its regular pattern of changing brightness as a function of time.  
Found by: WebCrawler  
<http://antwrp.gsfc.nasa.gov/apod/lib/glossary.html>

**ASTRONOMY**  
In this page I've written something about. ASTRONOMY. Versión en Castellano. WE CAN ALWAYS LEARN SOMETHING NEW ! Astronomy is the science of heavenly...  
Found by: AltaVista  
<http://www.geocities.com/Yosemite/Trails/7812/e-astronomia.html>  
<http://chandra.bgsu.edu/~gcd/Spacetime6.html>  
Back to the course home page. Back to Table of Contents. Chapter 6 The Transition from Aristotle to Newton I. Introduction The purpose of this chapter is to describe, in broad outlines, the contributions to the theory of space and time which were..  
Found by: HotBot  
<http://chandra.bgsu.edu/~gcd/Spacetime6.html>

**Expanding Universe: Astronomy**  
520 : 525 - Astronomy and Allied Sciences. 520 522 523 525. General astronomy sites, periodicals, associations, planetariums Observatories,...  
Found by: AltaVista  
<http://www.mtl.toronto.on.ca/centres/bsd/astromy/520-25.HTM>

**The Atheism of Astronomy**  
Library: Historical Documents: Woolsey Teller: Atheism Of Astronomy The Atheism Of Astronomy: A Refutation Of The Theory That The Universe Is Governed

Fig. 9



(19) World Intellectual Property Organization  
International Bureau(43) International Publication Date  
19 October 2000 (19.10.2000)

PCT

(10) International Publication Number  
WO 00/062198 A3(51) International Patent Classification<sup>7</sup>: G06F 17/30

(21) International Application Number: PCT/US00/10003

(22) International Filing Date: 13 April 2000 (13.04.2000)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:  
60/129,103 13 April 1999 (13.04.1999) US

(71) Applicant: INDRAWEB.COM, INC. [US/US]; P.O. Box 323, Concord, MA 01742 (US).

(72) Inventors: KON, Henry; 91 Bow Street, Lexington, MA 02420 (US). BURCH, George; P.O. Box 323, Concord, MA 01742 (US).

(74) Agents: KELLY, Edward, J. et al.; Foley, Hoag &amp; Eliot, LLP, One Post Office Square, Boston, MA 02109 (US).

(81) Designated States (national): AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CR, CU, CZ, DE, DK,

DM, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZW.

(84) Designated States (regional): ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:

— with international search report

(88) Date of publication of the international search report:  
26 September 2002

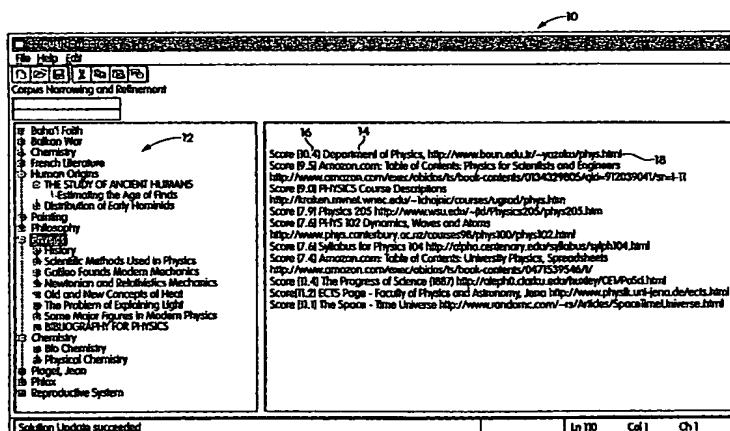
(15) Information about Correction:

Previous Correction:

see PCT Gazette No. 48/2001 of 29 November 2001, Section II

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: SYSTEMS AND METHODS FOR EMPLOYING AN ORTHOGONAL CORPUS FOR DOCUMENT INDEXING



(57) Abstract: The invention provides for indexing and cataloging of content on the Internet, as well as from other stores of information, may be performed by applying a process that employs an orthogonal corpus, or corpora, of information, such as an Encyclopedia. To this end, the processes described herein identify the topics discussed within the corpus. The process also identifies within the corpus a set of keywords that are relevant to the topics presented in the corpus. The keywords associated with a topic may be employed to identify documents stored in another database that are related to the topic. A graphical representation of the index of topics found in the corpus may then be generated, with individual topics operating as links to these related documents. Thus, a user interested in reviewing content in the corpus related to a certain topic, may also activate a link in the graphical representation of the index to access other documents that have been identified as related to the topic of interest to the user.

## INTERNATIONAL SEARCH REPORT

International Application No

PCT/US 00/10003

A. CLASSIFICATION OF SUBJECT MATTER  
IPC 7 G06F17/30

According to International Patent Classification (IPC) or to both national classification and IPC

## B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)  
IPC 7 G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

EPO-Internal, WPI Data, PAJ, INSPEC, IBM-TDB

## C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	<p>CHANG C-H ET AL: "Customizable multi-engine search tool with clustering" COMPUTER NETWORKS AND ISDN SYSTEMS, NORTH HOLLAND PUBLISHING, vol. 29, no. 8-13, 1 September 1997 (1997-09-01), pages 1217-1224, XP004095318 AMSTERDAM, NL ISSN: 0169-7552</p> <p>page 1218, right-hand column, line 19 -page 1219, left-hand column, line 15; figure 1 page 1220, right-hand column, line 2 - line 17 page 1221, left-hand column, line 1 - line 9; figure 4 page 1221, left-hand column, line 28 -right-hand column, line 20; figure 5</p> <p style="text-align: center;">-/-</p>	1-4, 9, 10, 13-19

☒ Further documents are listed in the continuation of box C.

☐ Patent family members are listed in annex.

## \* Special categories of cited documents:

- \*A\* document defining the general state of the art which is not considered to be of particular relevance
- \*E\* earlier document but published on or after the international filing date
- \*L\* document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
- \*O\* document referring to an oral disclosure, use, exhibition or other means
- \*P\* document published prior to the international filing date but later than the priority date claimed

\*T\* later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

\*X\* document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

\*Y\* document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

\*&\* document member of the same patent family

Date of the actual completion of the international search

17 January 2002

Date of mailing of the international search report

04/02/2002

Name and mailing address of the ISA

European Patent Office, P.B. 5818 Patentlaan 2  
NL - 2280 HV Rijswijk  
Tel. (+31-70) 340-2040, Tx. 31 651 epo nl,  
Fax: (+31-70) 340-3016

Authorized officer

Polzer, A

## INTERNATIONAL SEARCH REPORT

International Application No

PCT/US 00/10003

## C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	O'KANE K C: "Generating hierarchical document indices from common denominators in large document collections" INFORMATION PROCESSING & MANAGEMENT, ELSEVIER, vol. 32, no. 1, 1996, pages 105-115, XP004001098 BARKING, UK ISSN: 0306-4573	19,20
A	page 106, line 17 - line 35 page 108, line 15 -page 111, line 8; figures 4,5 page 114, line 7 - line 23; figures 7,8 ---	1
X	CHEN H ET AL: "Internet Categorization and Search: A Self-Organizing Approach" JOURNAL OF VISUAL COMMUNICATION AND IMAGE REPRESENTATION, ACADEMIC PRESS, INC., vol. 7, no. 1, 1 March 1996 (1996-03-01), pages 88-102, XP000619822 ISSN: 1047-3203	19
A	page 92, left-hand column, line 20 -right-hand column, line 2 page 95, left-hand column, line 31 - line 46; figures 2-4 ---	1,20
X	SAHAMI M ET AL: "SONIA: a service for organizing networked information autonomously" PROCEEDINGS OF DIGITAL LIBRARIES '98, THIRD ACM CONFERENCE ON DIGITAL LIBRARIES, PITTSBURGH, PA, US, 23 - 26 June 1998, pages 200-209, XP002187580 New York, NY, US ISBN: 0-89791-965-3	19
A	page 202, left-hand column, line 2 -right-hand column, line 17; figure 2 page 203, left-hand column, line 15 -right-hand column, line 5 -----	1,20

